DOCUMENT RESUME

ED 426 058 TM 028 270

TITLE	Research in the Schools, 1994-1998.
INSTITUTION	Mid-South Educational Research Association, MS.; Alabama Univ., Tuscaloosa.
ISSN	ISSN-1085-5300
PUB DATE	1998-00-00
NOTE	767p.; Published twice a year. Indicidual articles are covered in CIJE.
AVAILABLE FROM	Research in the Schools, University of Alabama at Birmingham, School of Education, 233 Educ. Bldg., 901 13th Street, South, Birmingham, AL 35294-1250 (Individual subscription, $25 per year; institutional subscription, $30 per year; foreign surcharge, $25 per year for individual and institutional).
PUB TYPE	Collected Works - Serials (022)
JOURNAL CIT	Research in the Schools; v1-5 1994-1998;
EDRS PRICE	MF04/PC31 Plus Postage.
DESCRIPTORS	*Academic Achievement; Educational Practices; *Educational Research; *Elementary Secondary Education; Instructional Leadership; Intelligence Tests; Multivariate Analysis; *Research Methodology; Scholarly Journals; Teaching Methods

ABSTRACT

This document consists of the first five items (10 issues) of the serial "Research in the Schools," a nationally refereed journal published by the Mid-South Educational Research Association and the University of Alabama. It publishes original contributions related to research in practice, topical issues in education, methods and techniques, assessment, and other topics of interest to educational researchers. Research in any area of education and involving any age group may be published. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 426 058

Research in the Schools

(ISSN-1085-5300)

Volumes 1-5, 1994-1998
(10 Issues)

TM 028 270

BEST COPY AVAILABLE

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and The University of Alabama.

Volume 1, Number 1 **Spring 1994**

Sixty Years of Research in the Schools: A Conversation with Ralph W. Tyler 1
James E. McLean

Small is Far Better 9
B. A. Nye, Charles M. Achilles, J. Boyd-Zaharisas, B. D. Fulton, and M. P. Wallenhorst

Aspirations of Minority High School Seniors in Relation to Health Professions Career Objectives 21
William A. Thomson, Leslie Michel Miller, Bernice Ochoa Shargey, James P. Denk, and Bruce Thompson

Leadership for Productive Schools 29
William L. Johnson, Karolyn J. Snyder, and Annabel M. Johnson

The Relationship of the Murphy-Meisgeier Type Indicator for Children to Sex, Race, and Fluid-Crystallized Intelligence on the KAIT at Ages 11 to 15 37
Alan S. Kaufman and James E. McLean

A Comparison of Two Procedures, the Mahalanobis Distance and the Andrews Pregibon Statistic, for Identifying Multivariate Outliers 49
Michele Glankler Jarrell

Gender Differences in Achievement Scores on the Metropolitan Achievement Test-6 and the Stanford Achievement Test-8 59
John R. Slate, Craig H. Jones, Rose Turnbough, and Lynn Bauschlicher

The Global Coherence Context in Educational Practice: A Comparison of Piecemeal and Whole-Theme Approaches to Learning and Teaching 63
Asghar Iran-Nejad

TM028270

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. A. Boser
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies; 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles; 3) *Methods and Techniques*--descriptions of innovative teaching strategies in research/measurement/statistics, descriptions of technology applications in the classroom, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses; 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like; and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (3rd ed., 1983), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, Office of Research and Service, The University of Alabama, P. O. Box 870231, Tuscaloosa, AL 35487-0231**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1994 by the Mid-South Educational Research Association.

ISSN 1085-5300

RESEARCH IN THE SCHOOLS

Volume 1, Number 1 **Spring 1994**

EDITORS
James E. McLean and Alan S. Kaufman, *The University of Alabama*

PRODUCTION EDITOR
Margaret L. Glowacki, *The University of Alabama*

EDITORIAL ASSISTANT
Anna Williams, *The University of Alabama*

EDITORIAL BOARD
Charles M. Achilles, *University of North Carolina at Greensboro*
Mark Baron, *University of South Dakota*
Michèle Carlier, *University of Reims Champagne Ardenne (France)*
Sheldon B. Clark, *Oak Ridge Institute for Science and Education*
Michael Courtney, *Henry Clay High School (Lexington, KY)*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Kentucky*
Donald F. DeMoulin, *University of Missouri--Columbia*
R. Tony Eichelberger, *University of Pittsburg*
Daniel Fasko, Jr., *Morehead State University*
Patrick Ferguson, *Arkansas Tech University*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Samuel Hinton, *Eastern Kentucky University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
Randy W. Kamphaus, *University of Georgia*
Jwa K. Kim, Middle *Tennessee State University*
Jimmy D. Lindsey, *Southern University and A & M College*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Roy P. Martin, *University of Georgia*
Peter Melchers, *University of Cologne (Germany)*
Claire Meljac, *Psychologue AU C.H.S. Sainte-Anne (France)*
Soo-Back Moon, *Hyosung Women's University (Korea)*
Arnold J. Moore, *Mississippi State University*
Thomas D. Oakland, *University of Texas*
William W. Purkey, *University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Clemson University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Arkansas State University*
Bruce Thompson, *Texas A & M University*
Anne G. Tishler, *The University of Montevallo*
Wayne J. Urban, *Georgia State University*

GRADUATE STUDENT EDITORIAL BOARD
Vicki Benson, *The University of Alabama*
Ann T. Georgian, *The University of Southern Mississippi*
Jin-Gyu Kim, *The University of Alabama*
Robert T. Marsousky, *University of South Alabama*
Jerry G. Mathews, *Mississippi State University*
Dawn Ossont, *Auburn University*
Malenna A. Sumrall, *The University of Alabama*
Carol J. Templeton, *Mississippi State University*

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 1-8

Sixty Years of Research in the Schools: A Conversation with Ralph W. Tyler

conducted by
James E. McLean

For generations, students in education have been influenced by Ralph Tyler's work. His influence has spanned over 60 years. I first encountered Dr. Tyler's work in undergraduate school, and his influence on my graduate education, during the early 1970s, was formidable. I remember being impressed with how much he was able to say in a few words and how clearly he was able to say it. My first face-to-face meeting with Dr. Tyler was at an annual meeting of the American Educational Research Association (AERA) in the early 1970s. Having read many of his works, I was even more impressed with him in person. It seemed not to matter to him that I was a graduate student at the time. I remember being both impressed and flattered that he treated me as a full-fledged colleague. This was a highlight of my doctoral career.

Since that time, we have maintained our professional relationship, and I came to respect him even more as a person than I did for his considerable contributions to education. We would renew our acquaintance often at professional meetings such as AERA and the American Evaluation Association. He was a frequent visitor to The University of Alabama campus. Faculty and students were fortunate to have Dr. Tyler as a visiting professor at Alabama in the spring of 1979. During the time I have known Dr. Tyler personally, he has never failed to respond to a request. Over the years, he has made presentations to my classes, helped me think through research problems, served at my request as a keynote speaker for a Mid-South Educational Research Association annual meeting, and agreed to this interview for *RESEARCH IN THE SCHOOLS*. He first was going to record an article for the journal but, when his health precluded that, he wrote offering to pay my way to his Milpitas, California home to conduct this interview. That is the kind of man Dr. Tyler is. Personal financial gain has never been one of his motives. He never accepted even a small honorarium for the activities I noted above.

Dr. Tyler's influence on education is still very much with us today. Many current practices in education had their roots with Dr. Tyler. These practices range from curriculum development to evaluation. Dr. Tyler was the first to apply the terms "evaluation" and "assessment" in the educational arena. This has resulted in his being referred to as the "father of educational evaluation." Also among his or his students' contributions are objective-based curriculum, achievement testing, criterion-referenced testing, item-banking, objective-based evaluation, mastery learning, and the taxonomic classification of educational objectives. He also was instrumental in the development and implementation of the federal Regional Laboratories, the National Assessment of Educational Progress (NAEP), and the American College Testing program. Perhaps his most influential contribution was through his book, first published in 1950, *Basic Principles of Curriculum and Instruction*. This book clearly expresses Dr. Tyler's philosophy that curriculum and evaluation are inseparable.

The interview that follows took place in Dr. Tyler's home in Milpitas, California, during two beautiful days in May 1993. Dr. Tyler (born April 22, 1902) had just celebrated his 91st birthday at the time of the interview. While his eyesight and hearing are not what they once were, his mind was clear. This was demonstrated by his recall of detailed information from as early as the 1920s. Dr. Tyler now lives with family in San Diego, California. He can be reached at 701 Kettner Boulevard #198, San Diego, CA 92101.

McLean: During the past 60 years, your work has had and continues to have a major influence on educators and educational practice. What would you say were the defining moments in your career?

James E. McLean is a University Research Professor at The University of Alabama and co-editor of *RESEARCH IN THE SCHOOLS*. Ralph W. Tyler is currently residing in San Diego, California. Recent health problems have curtailed his professional activities. Correspondence should be sent to Dr. Ralph W. Tyler, 701 Kettner Boulevard #198, San Diego, CA 92101. It is difficult for him to reply but he would still enjoy hearing from his friends.

Tyler: Of course my family was the first influence. I was born in Chicago where my father, who was a physician, was attending theological seminary after deciding to become a minister. I grew up in Nebraska, got my bachelor's degree at Doane College, a small Congregational college in Crete, Nebraska. Then I went to Pierre, South Dakota, to teach. I became deeply interested in teaching the varied types of students we had. We had Indians from the reservations; we had migrant farm workers who came up from Mexico to work in the sugar beet fields; we had children of European immigrants; we had the children of drifters who were around the railroad tracks which were near the school. I discovered it was so interesting to work with them. For example, the first thing that happened when I went to my first class was that two American Indians came up to me and said, "We're gonna beat you up." Well, I said, "You *can* beat me up, that'll take only one of you to do it. What do you want to beat me up for?" "Well, you make us go to school. We don't like to go to school." "I don't make you. Why do you go to school if you don't like it?" "Well, we want to play football." "If you play football, you've got to be eligible, haven't you?" "Yes." "To be eligible, you've got to pass this course?" "Yes." "Well, what are you gonna do about it?" "I don't know." "Why don't you come with me and we'll go into the laboratory, get out the equipment, and do some experiments?" That challenged them to get interested in their learning instead of trying to avoid it. That also started me on my career, and I've never wanted to be anything but a teacher since. By the way, I ran into the grandson of one of those Indian boys at a recent AERA meeting. He told me that I so excited those kids that they became teachers themselves, and he was a teacher, too.

McLean: This kind of story illustrates why teaching is so rewarding. You were fortunate to get the feedback from the grandson, as most of us don't hear about our successes. Where did this job lead you?

Tyler: Then I went to the University of Nebraska to be a supervisor of practice teaching. I also got a master's degree and then went to Chicago to get my doctorate. When I got my doctorate in 1927, my first job was a professorship at the University of North Carolina at Chapel Hill. My job was to work with the schools of the state by helping them with their curriculum problems. So I spent one day a week, on Mondays, with my graduate students in Chapel Hill and went out into the field the rest of the week. I drove down to Wilmington and started back, working first with the schools in New Hanover County, then moving on up the state until I finally got back home on Friday evening. By working with the schools that way, seeing what they were teaching and helping them teach more effectively, I became greatly enthralled, and that only increased my commitment to teaching.

McLean: How long did you remain in North Carolina? Did you go to Ohio State from there?

Tyler: Well, I was at North Carolina in 1927, 1928, and 1929. At that time, I was invited to come to Ohio State to work in the Bureau of Educational Research with W. W. Charters, whom I had worked with at Chicago and had greatly admired. So I took the job with him as head of the Division of Accomplishment Testing in the Bureau of Educational Research at Ohio State. There I worked with the schools of the state and with faculty from the University itself on problems that they had. I discovered a number of things. One was that they didn't solve problems by talking about them; it required doing something that modified the problem. That didn't start from the top down; you had to start from the bottom up. So I worked with them for nine years. Then I was invited to come to the University of Chicago to take the place of my mentor, Charles Judd. I eagerly did that and came to Chicago in 1938. I was there until I was invited to become head of the Center for Advanced Study in the Behavioral Sciences in 1955. So I was at the University of Chicago working with the faculty on problems there from 1938 to 1955.

McLean: The Center for Advanced Study in the Behavioral Sciences was at Stanford--is that when you moved to this area?

Tyler: That's right. And I didn't want to go back after I had once been out here.

McLean: I can understand that. The weather here and this area certainly are beautiful. You told us what inspired you to become a teacher and provided us with a wonderful thumb-nail sketch of your career. Who are some of the people that most influenced you professionally?

Tyler: I read a lot, so it's hard to say exactly where many of the ideas came from. People who influenced me included C. H. Judd and W. W. Charters at Chicago. I admired the work of John Dewey and Edward Thorndike. Of course, there were many others, among them my students.

McLean: I know that you have had many students over the years who went on to make names for themselves in educational research. Would you tell us about a few of these?

Tyler: Well, we know John Goodlad, who came to the University of Chicago from a principalship in Vancouver or somewhere in British Columbia and was going to work on the disadvantaged emotionally disturbed children. He got so interested in his curriculum course that he changed majors. Then there was Hilda Taba who came from Estonia. She came here originally to get a master's in philosophy from Bryn Mawr and then the Russians took over Estonia and she didn't want to go back. She got caught up with pretty soon by the Immigration Service because she came on a student visa, so I arranged for her to go back to Canada and apply for a resident visa and then I sponsored her coming in that way. Then there was Herb Thelen, who was working for a degree in chemistry and got interested in understanding human behavior, and he decided to change his major from chemistry to curriculum and development. Ben Bloom was my graduate assistant at Chicago and later worked with me there. When I left the University of Chicago, he took over the responsibility of the Evaluation Center at the University of Chicago. A very bright young man. He was planning to major in statistics, but he found curriculum so interesting that he changed his mind. Also, there was Chester Harris, Christine McGuire, Lee Chronbach, and Bruno Bettelheim. I can't leave out Ray Loree. He came from Manitoba, Canada, to get his doctorate at Chicago. A very bright and able man. I know he worked with you at Alabama.

McLean: Actually, he hired me at Alabama. Probably because I told him I admired your work.

Tyler: Well, there are many others. These are just illustrations.

McLean: As you said, you were committed to teaching from an early age. What influenced you to leave the classroom and become an educational researcher?

Tyler: I heard so many claims about education's failure that I knew weren't true, because I had spent too much time in the schools. Until I got ill, I spent an average of about 10 hours a week in schools. What was said about them, for example in *A Nation at Risk*, was absolutely untrue, and I felt it was time that educational researchers really looked at the problems and found out what the facts really were.

McLean: Do you think that educational researchers should spend more time in the schools, then?

Tyler: Yes. A science is built upon careful observation and repeated study of the situation. Researchers that try to draw conclusions without getting down there don't know what they're doing.

McLean: You have already mentioned the *Nation at Risk* report that came out about 10 years ago. One thing this report did do, whether it was accurate or not accurate, was to stimulate the reform movement of the last 10 years.

Tyler: Movements that are not based on facts don't get very far. They pass like fads. You start with the problem in a particular situation. You study it carefully and make repeated observations, then begin to understand it. You check that out, especially getting the cooperation of teachers and parents who are also working with children. This careful observation and development was never done in *A Nation at Risk*. It was shooting from the hip on the basis of impressions.

McLean: So you think that reform should be based more on local problems than on some global ideas?

Tyler: Education is a distinctly family affair. It is not a statistical one where the elements make little difference. It's *your* child that we're concerned with. It's *your* teacher we're concerned with. We've got to understand that well enough to figure out what's the problem they're facing.

McLean: While we are talking about various reform movements, many are going on today and even though you weren't very complimentary on what started them, these movements seem to have many common recommendations. What is your opinion of non-graded classrooms?

Tyler: There are some basic principles of learning. Begin where the student is and move step by step, but whether non-graded classes would help depends on how they are organized.

McLean: Is your answer the same for multi-age grouping?

Tyler: Yes. The ideas are good if they are implemented with an understanding of child development. You know, whenever there is a new reform movement people want to get on the bandwagon. They say they're doing it when they really haven't changed at all.

McLean: One of the things that is different about this reform movement is that everyone seems to involve the use of modern computer technology. Do you feel computer technology has the potential to assist educators?

Tyler: Of course, computers have been a big help to educational researchers for data analysis and handling data. Educational technology can also be an aid to teachers. But, only if the computer is used for more than drill and practice. Educational researchers have to find ways that technology can help the teacher. To do that, researchers must spend time in the classroom. My general experience has been that technology has interfered with the learning of students rather than helping them. To avoid these problems in the future, researchers must find out what teachers' problems are before they can design programs to help them.

McLean: One of the major uses of technology right now is based on its ability to store and retrieve information. In other words, to provide immediate access to the worldwide storehouse of knowledge so students won't have to memorize it.

Tyler: The learning of children is not just generalized knowledge. It's the process by which a child learned that information. Just putting the information on a computer does not solve the problem. You have to find ways the child can use this information to help the child become a better person. What attitudes does the child have toward the world? You also have to help the child develop enthusiasm and a fine outlook. If a child feels, "Oh, this is a terrible life. We've got to go to work again today," a lot of knowledge will not help. Everything depends so much on the child's attitude--that's not generalized information. What I am saying is that even if computers reach their potential, that is not enough. The childrens' attitudes are also important.

McLean: Do you think that having information available to students, so that they won't have to memorize it, will allow teachers more time to work on problem solving abilities?

Tyler: Well, they certainly shouldn't be memorizing facts, except if they are needed. What's the other possibility? To work on their own problems like the Paideia schools, as was recommended by Mortimer Adler in *The Paideia Proposal* [Macmillan Publishing Company, Inc., New York, 1982]. In the Paideia schools, they begin with the problems of the student, not with the subject matter. Subject matter is derived from knowledge you discover you need in order to solve your problems.

McLean: So the schools should be centered around the needs of the students--not for the convenience of teachers or the administrators?

Tyler: Well, what are schools for? They are not for the teachers, but for the children!

McLean: There are many curriculum researchers and educational researchers, including yourself and John Dewey, who have long said that the greatest source of knowledge for improving education is in the teachers themselves. Do you think that's the direction educational research should go?

Tyler: Research should be based first on the problem. What's the problem you're working on? If the problem is one of early childhood education, you want to know more about the family upbringing, how the child was treated early. If the problem is a high school level one, you may want to understand more about the teachers' backgrounds and what they believe and try to do. You get to the root of a problem by observing classrooms and talking with teachers.

McLean: Do you think that teachers themselves should be more involved in the research process?

Tyler: They certainly should be, along with parents. The problems can't be understood unless you see them from the point of view of the student, the teacher, and the parent. When I worked with the Coalition for School Improvement in Massachusetts, we had a committee of teachers, principals, and parents continually identifying problems and trying to understand them. So to try to study something outside of who does it is intruding, and you won't understand what they're doing. Most of human behavior is operated from purpose, not from simple casual relationships.

McLean: One of the major criticisms of educational research is that its findings have not influenced educational practice. Maybe if research were based on specific educational problems, that criticism would no longer be valid.

Tyler: Exactly. Most research in other fields goes back to problems that are interfering with their effectiveness, and if the teachers' problems are understood, they would want to do something about them. It's *their* problems and *they* are concerned with doing a better job.

McLean: The National Council for the Accreditation of Teacher Education or NCATE now requires a college to have what they call a knowledge base to become accredited. In other words, their programs have to be based on some theory or have some theoretical basis. What would you think should be the theoretical basis for the programs in a college of education?

Tyler: Try to understand the problems of their students. I don't think that you start with the theory. You start, as other scientists do, with observations of difficulties or problems, so there's no knowledge base until you identify the knowledge as you work with the problems. What is the knowledge base of the problems that you face? You can't define that in advance. It comes from the effort to understand your experience. As Dewey said, knowledge comes from experience. Whitehead said, "Knowledge is like fish; it won't keep." Knowledge is something that comes from understanding experience and is not understood before you have the experience.

McLean: Do you think that teachers should learn about so-called action research as part of their programs so that they can be their own problem solvers?

Tyler: As part of their curriculum, they should be working with the problems of learning. One of the things recommended by the National Commission on Teacher Education was that as they finish high school they would begin to work on problems in informal education agencies, such as 4-H Clubs, Girl Scouts, things of that sort, and as they work on those problems and hold seminars once a week or once every two weeks to discuss those problems, they will understand the theory, because theory is derived from practice and not practice from theory.

McLean: The field of educational measurement is undergoing a good bit of change recently. The last annual meeting of the American Educational

Research Association and the National Council for Measurement in Education featured numerous sessions on so-called performance assessment, or what they're now calling authentic assessment. My limited knowledge of measurement history suggests these things are not new. Do you recall the use of performance assessment in past years, and do you think performance assessment can be useful today?

Tyler: When you consider any assessment, the first few questions are: For what purpose? Who wants the data? How will they use it? In what form should it develop? The next question is: How can we get this information? The notion that standardized tests alone will give it to us is not true. Standardized tests only tell you about what is being measured by the test, and what you want to know is what will help you understand the teachers' problems. So we ought to reduce our dependence on standardized tests. Go back to what was called performance assessment, namely finding out whether we are really accomplishing what we want to accomplish from practical situations.

McLean: Much of this work with standardized tests has been government supported and conducted by large testing companies like Educational Testing Service and, more recently, by American College Testing or ACT. One example of this is the National Assessment of Educational Progress or NAEP. I know that you have influenced both testing companies and NAEP. I think you helped get the ACT and NAEP started. What do you think these and other national testing programs should be doing today?

Tyler: They should also be examining how the tests they are developing are being used. Standard tests need to be developed for a particular purpose and then used for *that* purpose only. What they often do is bring before teachers problems of education, rather than starting with the problems. Standard tests need to help teachers with instruction, not just demonstrate problems teachers can't control. More time needs to be spent letting teachers know how to use the information they get from standardized tests.

McLean: The *Appraising and Recording Student Progress* study that you evaluated in the 1930s that's commonly called the *Eight-Year Study* still influences curriculum today. Could you tell us a little about the *Eight-Year Study*? What was its purpose, and what do you think it accomplished?

Tyler: In 1931 there was great criticism of college admission requirements that began specifying the courses one should take and the number of hours of credit one must have to be admitted to some prestigious colleges. Some young Turks of the time challenged these requirements as being inappropriate solutions to the colleges' problems--saying they ought to begin with the problems and not with the "solution." Finally, it was agreed by the College Entrance Board and by the state departments of education involved that for several years the candidates for college admission would be freed from meeting these new requirements and to let them work out requirements that would meet their needs. Well, at the end of the first year the students admitted under this program found they were going to be measured by the same standard of achievement as the other students, and they rebelled. They said we're not going to go on with this experiment if we are going to be measured by something that doesn't represent what we're trying to do, and this impasse was solved when they suggested, then, that they could have different degree requirements for different students. The question became: What does this youngster need to do next in his educational career and his schooling career? Instead of setting up definite requirements as was suggested, they would be set up in terms of the students' needs. Well, the state departments didn't think that was a very good idea. A friend prevailed upon the state departments of education and the College Entrance Board to let a selected group of schools demonstrate what they could do if they based their programs on the needs of their students, rather than on the new requirements for college entrance that were specified by the state departments of education. This was agreed to, and the principals began to specify what the students should have. Then there began new complaints, because the principals didn't know any more than the College Board about what their students were like. I suggested that we spend the time in working with the students to find out what they needed to try to get the plans for their future worked out in terms of their needs rather than in terms of the requirements of the College Board, the state departments of education, or the colleges themselves. So that started the *Eight-Year Study*. They were given eight years to demonstrate that schools could work out their own programs successfully. A control group of schools that didn't operate that way was also set up. At the end of the period it was found that the experimental group was superior on most of the things they did--they did not fall down because they didn't meet the College Board requirements. When they worked with the students' own needs, they were able to get students who were more successful than those students who followed the College Board requirements or the state department requirements.

McLean: I understand that this study served as a training ground for a whole generation of educational researchers. Can you tell us a few of the people who worked on this study with you?

Tyler: Yes! Over the eight years of the study, I had three associate directors of evaluation--Oscar Buros, Louis Raths, and Maurice Hartung. Some of my other associates and assistants were Bruno Bettelheim, Hilda Taba, Harold Trimble, Christine McGuire, and Chester Harris. There were many others, but I cannot recall all their names right now.

McLean: You've had a good bit of overseas experience, too. You've consulted with the governments in a lot of countries concerning their educational systems including Russia and China. How were you able to help them improve their educational systems?

Tyler: Understand--I'm not necessarily helping them if they didn't want to be helped. For example, Russia didn't want to be helped, but I was interested in learning how they were operating. That was in 1961. You're right. I've worked with a number of overseas countries trying to understand their problems and, as far as possible, help them. But helping them depends on their purpose. To help China, which was trying to become more totalitarian, or any totalitarian country, was not my purpose. You've got to believe in what you are trying to do. It is not appropriate to force kids to believe certain things. Indoctrinate them. So that, to be able to help them, would help them indoctrinate children, and I don't believe in that. I could understand it, but I didn't believe in it. On the other hand, to help them understand what they're doing, as in some of the black countries in South Africa, is very helpful. I've learned a good deal by working in other countries. I've worked in every continent except Australia. I've never had the opportunity to work in Australia.

McLean: From your previous comments, I would guess that the first thing you do when you go to work with another country is to visit their schools and learn about their problems.

Tyler: The first opportunity I had to go to another country was in 1967 when I was invited to deliver a series of lectures in Israel. I went there and gave my lectures on what the curriculum was about. This so interested them that they asked me to come back and to help them work on democratizing their own classes. So I went back and forth to Israel at least six times. In Indonesia, they wanted me to help them get over the problems that they had as a result of the dictatorship of the Communist Party and when General Suharto was trying to make them become Fascists. Working in Tanzania, where practically everybody is a farmer and their literacy rate is very low, I helped them work out family learning so farmers could choose the times they wanted to work in the fields. Whenever the family wanted to work together, the parents would learn and the children would learn. So using available resources to meet educational problems became something that interested me a good deal.

McLean: That sounds fascinating. I know from working with foreign students at the University that the ones who come for graduate study obviously are some of the best, but I've observed their learning styles vary widely, often based on their own educational systems.

Tyler: Yes. The aims of education should be the same in America--trying to help students become responsible democratic citizens. But the way they get there depends a good deal upon their previous experience and how they've learned to control their own learning. Don't try to make them like everybody else. One of the dangers is to say, "This is the way you should learn" instead of saying, "Let's help you discover the ways you can learn best and help you promote them."

McLean: You know that one of the reasons I got this opportunity to meet with you is because of the journal the Mid-South Educational Research Association is starting, a national journal called *RESEARCH IN THE SCHOOLS*. We chose that name to put the emphasis on research that could help improve schools. What advice would you have for Alan Kaufman and me as editors of this journal?

Tyler: To select your articles and papers based on problems that really exist, not to depend on papers written at a desk at home, but to go out and work with schools, get to understand the situations there thoroughly, and begin to say what the problems really are. The tendency of journals is to begin from the top down instead of from the bottom up.

McLean: That is good advice for any educational researcher. Thank you for your time and for your many insightful comments.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 9-20

Small Is Far Better

B. A. Nye, C. M. Achilles, J. Boyd-Zaharias, B. D. Fulton, M. P. Wallenhorst

The Lasting Benefits Study (LBS) is following up the pervasive effects of small classes for primary-grade students in Tennessee's Student/Teacher Achievement Ratio (STAR) Project. Project STAR, a randomized, longitudinal, statewide experiment, demonstrated that students in small classes (15:1) had statistically significant and educationally significant advantages over students in other classes. Students in STAR classes for at least the third grade participated in LBS. MANOVA analysis for unequal n's revealed that statistically significant ($p \leq .01$ or better) achievement benefits from participation in small K-3 classes remained after students returned to regular-size fourth and fifth grade classes. Results were consistent for all measures across all locations. Project Challenge extends class-size results more widely as a policy initiative.

Introduction: Class-Size Issues in a New Dimension

Educators have debated the issue of class size for years. Bloom (1984) posed the "2-sigma" problem, asking how education and society could find an affordable way to attain in some group setting the pupil achievement attained in one-on-one tutoring. Bloom's and other research (e.g., Slavin, 1989, 1990) focused the idea of small class size benefits on achievement, but class-size research is expensive and time consuming.

Part of education's problem is to address the needs of those whom education is asked to serve. For public education, these are the young people who enter the schools. The comfortable former assumption of schooling (two middle-class biological parents in the home with one parent working) does not hold up today. Hodgkinson (1991) states new demographic realities:

> Since 1987, one-fourth of all preschool children in the United States have been in poverty. . . . This is the nature of education's leaky roof: about one-third of preschool children are destined for school failure because of poverty, neglect, sickness, handicapping conditions. . . 23% of America's smallest children (birth to age 5) live in poverty, the highest rate of any industrialized nation. (pp. 10-11)

In today's schools, incoming students are increasingly hindered by poverty, parental drug/alcohol use, and by effects of low birth-weight (a frequent partner of teen pregnancy and no prenatal care). Educators must make adjustments--at least in the early primary grades--to accommodate changing clients and client needs. Hamburg (1992) makes a strong link between childhood health and the possibility of a pupil benefitting from education. "A recurrent theme . . . is the close relationship of education and health. Children in poor health have difficulty in learning" (p. 84). News media daily report on homelessness and changing family structure (one-parent settings, both parents working, etc.). Hamburg addresses the impact of family stability on early childhood development. "Families can be disrupted in a variety of ways--through poverty, social disadvantage . . . homelessness--that in turn challenge a child's natural development" (p. 98).

Consider the burden that these problems place on teachers who work with these children in their first years in schools. Years ago, when fewer school entrants were from impoverished or disrupted families, teachers might have been able to work effectively with 30 or more pupils. Some school leaders countered early demographic changes by making teacher aides available to work with one or more teachers. Another alternative is to

---

This article was based on a report on three class-size initiatives: Tennessee's Student Teacher Achievement Ratio (STAR) Project (8/85-8/89), Lasting Benefits Study (LBS: 9/89-7/93), and Project Challenge (7/89-7/93) as a Policy Application (Preliminary Results). The paper, originally presented at the 1992 Annual Meeting of the Mid-South Educational Research Association won that organization's 1992 Outstanding Research Paper Award. The authors acknowledge the contributions of the entire Student Teacher Achievement Ratio (STAR) Project staff, especially to E. Word, Tennessee State Department of Education, Project Director; H. Bain, J. Folger, J. Johnston, and N. Lintz who were the other members of the STAR Consortium; J. Finn, R. Hooper, and G. Bobbett, Consultants. B. A. Nye, Director, Lasting Benefits Study and Center of Excellence for Research in Basic Skills, Tennessee State University, Nashville, TN 37203. C. M. Achilles, Professor, Education Administration, UNC-Greensboro, Greensboro, NC 27412-5001. Member of the Star Consortium 8/85-9/88, and consultant to LBS, 89-94. J. Boyd-Zaharias, B. D. Fulton, and M. P. Wallenhorst, Staff of LBS at the Center for Excellence.

have fairly small classes for all pupils, especially in early grades--a change from "assembly-line" to "case-load" approaches. There *are* small classes for *special* students (e.g., handicapped, vocational, gifted). Aren't all pupils special? Aren't the new entrants to schooling who come from disadvantaged backgrounds special? Interestingly, the area of small-class benefits to pupils has been quite thoroughly researched. Yet, policy makers hesitate to use the evident solution. While they dally trying to find better (and cheaper) alternatives the conditions worsen, especially, as Hamburg (1992) says, for *Today's Children*. Perhaps, like the fabled tortoise and hare, the consistent tortoise of class-size results may plod into the lead.

Education researchers seldom conduct either experimental or longitudinal studies. Education research does not often provide clear direction for education practice. In contrast, this paper presents a continuing strand of research that (a) began in 1985 as experimental and longitudinal (through 1989), (b) is still using and extending the original data base (1989-1994), (c) has provided policy direction *and* implementation (1989-1994), and (d) is spawning a variety of interesting ancillary studies. Table 1 shows relationships of the studies. The discussion is divided into Phases I, II, and III.

Some things make so much sense that people wonder why researchers study them. Class size--the number of pupils that a teacher works with at a given time--is one such issue. Early studies were usually short-term, poorly designed, and dealt with reductions in large units (say 45-30 pupils). A controversial meta-analysis (Glass & Smith, 1978) and critiques of it (Education Research Service, or ERS, 1978, 1980) heated up the debate. Continuing policy discussions (Glass, Cahen, Smith, & Filby, 1982; Cahen, Filby, McCutcheon, & Kyle, 1983) encouraged Tennessee legislators to commission a large-scale, longitudinal experiment of class-size issues. While Tennessee's Student/Teacher Achievement Ratio (STAR) study was ongoing, policy debates continued (Mueller, Chase, & Walden, 1988; Tomlinson, 1988; Mitchell, Carson, & Badarak, 1989).

After STAR results became public (Word, Johnston, Bain, Fulton, Zaharias, Lintz, Achilles, Folger, & Breda, 1990), some collections of works on class size reviewed the findings and ideas related to policy (e.g., Robinson, 1990; *Contemporary Education*, 1990; *Peabody Journal of Education*, J. Folger (Ed.), 1989, published in 1992). The Robinson (1990) report did not yet have complete details from STAR, but did say, "Tennessee's Project STAR, currently in progress . . . had positive effects as measured by scores on nationally standardized tests (grades K-2)" (p. 82). Other studies reported generally positive results for STAR and mixed results for other "class size" studies. Tomlinson (1990) said, "Project STAR is doubtless the all time most comprehensive controlled examination of the thesis that a substantial reduction in class size will, of itself, improve achievement" (p. 19).

The Orlich (1991) statement is gratifying: "In my own opinion, (STAR) is the most significant educational research done in the US during the past 25 years" (p. 632). Two strong positive comments were: "This experiment yields unambiguous evidence of a significant class size effect, at least in the primary years" (Finn, Achilles, Bain, Folger, Johnston, Lintz, & Word, 1990, p. 135), and "This research leaves no doubt that small classes have an advantage over larger classes in reading and mathematics in the early primary grades" (Finn & Achilles, 1990, p. 573).

Table 1
Relationships of STAR, LBS, and Challenge
Showing Years Grades, Measurements, etc; 1985-1994

Study	Years	Grades	Measurement	Instruments
STAR*	1985-89	K - 3 1 grade/yr	Each year & longitudinal	SAT/BSF & questionnaires
LBS*	1990-94	4 - 8		
Cognitive	1990-93	4 - 7	Each year	TCAP
Particip.	1990, 1994	4	Grade 4	Questionnaire
Challenge**	1989-94	K - 3 Every year	Grade 2	TCAP

* Pupils progressed through the grades and were tested each year; by 1994 they are in grade 8.

** All pupils in grades K - 3 every year; tested in grade 2 only. LBS and Challenge are expected to continue.

Phase I. STAR:
The Basic Study and Database: Design and Scope

Project STAR began in 1985 with pupils in Kindergarten (K). All Tennessee districts were asked to participate. Due to the scope of the study, researchers (using a "power analysis") determined that they would need approximately 100 classes of each of three class types (S with average 1:15 teacher/pupil ratio--range 1:13-1:17; R with 1:24 average--1:22-1:26 range; and RA with 1:24 average and a full-time aide). Forty-two of the 140 districts (1985) were selected, and 79 elementary schools in those districts voluntarily provided the sites for STAR intervention. Three districts eventually dropped out.

Sites had to agree to participate for *four* years, to have some visitations and extra testing, and to allow random assignment of pupils and teachers to conditions. Sites had to have space for the added classes and at least 57 pupils in K. This did exclude very small schools from the study, but at least 57 pupils were needed for the in-school design (minimum of 1:13, 1:22, 1:22) that assured that any school with the S class also included R and RA class conditions. This powerful design helped ameliorate building-level variables such as leadership, curriculum, facilities, expenditures, SES, calendar, etc.

The state paid for additional teachers and aides for the four-year study (K-3) from 1985-1989. The STAR study made *only* class-size changes. Districts followed their own policies, curricula, etc. No pupil in STAR would receive less (e.g., would have a disadvantage from the state norm) by being in STAR. Not every pupil took every test or had every data point, so for a given year, the *n* for analysis was less than the total of pupils participating for that year. (Table 2 shows that 5,734 of the 6,325 K pupils provided the K analysis group.) *All pupils in an analysis had all data needed for that analysis.*

STAR employees monitored testing conditions for consistency. Although the pupil was the *primary unit* of data collection (researchers collected teacher, principal, and district data and such things as teacher interviews, etc., to support the class size analysis), the *class* was the *unit of analysis* (it was a study of class size effects). This analysis recognized that each pupil is *not* an independent measure--the teacher and classmates all influence the learning environment.

Legislation required that STAR classes be in four locations: inner city, urban, suburban, and rural. The major question was: "What is the effect of reduced class size (e.g., 1:15) on pupil achievement and development in K-3?" Research was conducted by a

Table 2
Parameters of STAR: Totals and Research Tapes, Grades K - 1

	Dist.	School	Pupils	Classes (N & %)							
				S		R		RA		Total	
1985-86 (K)	N	N	N	N	%	N	%	N	%	N	%
Totals	42	79	6325	127	38.7	103	31.4	98	29.9	328	100
Res Tape*	42	79	5734	127	38.7	103	31.4	98	29.9	328	100
1986-87 (Grade 1)											
Totals	42	76	7103	124	35.7	115	33.2	108	31.1	347	100
Res Tape*	42	76	5905	124	35.7	115	33.2	108	31.1	347	100

Note. S = 1:15; R = Regular; RA = Regular with Teacher Aide.

* The research tape included pupils who met various criteria. Not all pupils had scores for all measures each year. Participation in grade 1 is greater than in kindergarten (K) due to Tennessee not having required kindergarten (K); new pupils entered and were randomly assigned.

consortium of four universities, each with a principal investigator and staff (University of Tennessee, Memphis State, Tennessee State, and Vanderbilt) and the Tennessee State Education Agency (SEA) where the director was housed. Persons from each university monitored the study in assigned schools. (Ancillary studies reviewed training effects, teacher/teaching practices, etc.) *This* report *primarily* reviews achievement.

Achievement was determined by pupil scores on both Norm-Referenced Tests (NRT) and Criterion-Referenced Tests (CRT) appropriate for the grades. The CRT was Tennessee's Basic Skills First (BSF) test tied to the state curricula.

Due to the randomness, the basic design was posttest only (pretest in K was not an option). With scaled scores, it was possible to study year-to-year gains as STAR tracked *each pupil* and as pupils were in the same class size condition from year to year. When pupils moved to/from STAR schools, replacement was random.

*STAR Design/Analysis/Selected Findings**

The general multivariate design included four *locations* and the class type (S, R, RA) for either achievement measures or noncognitive measures. The design also included pupil (and teacher) characteristics of interest, and in grade 2 issues of teacher training. The primary analyses addressed the required questions as stated in the legislation and were completed for each of the four years. Additional longitudinal analyses are underway. (Details are available in STAR technical reports from the Center of Excellence, Tennessee State University, Nashville, TN 37203.) The outline for the primary analysis and the extended model for the detailed analyses appear elsewhere (e.g., Finn & Achilles, 1990; Word et al., 1990). The primary analysis consisted of mean differences between and among the groups being analyzed. [This design is also being followed in the Lasting Benefits Study (or LBS) to the degree possible.]

The analysis employed a general linear model approach for unequal-*n* design. The design has unequal *n*'s and some empty cells and requires multiple error terms to test all of the fixed effects. Test statistics were the univariate *F*-ratio for each measure and Wilks' likelihood ratio for multivariate sets. Other analyses and tests (e.g., chi square, correlation, regression) were employed as needed. There were two planned contrasts tested among three class types:

S class mean vs. all R and RA class means
(S vs. "Other")

R class mean vs. RA class mean

The major *achievement* results of STAR appear in Table 3. (For STAR, *development* measures such as attendance, discipline and self-concept showed no differences between S and R/RA.) In many ways, the monotony of the findings is significant. Essentially, pupils in S did statistically significantly better (usually at $p \leq .001$) than pupils in R and/or RA. *The class size effect was found equally in all locations (e.g., urban, rural) and favored the S condition in all four grade levels.* Less pervasive findings appeared in one or two grades.

Some simple analyses demonstrated powerful effects. Note (Table 4) that in the average percent of pupils passing the CRT (BSF) in grade 1 there appears to be a strong positive class size benefit for minority pupils. (This result was confirmed in more "sophisticated" analyses, but the results in Table 4 speak for themselves.) *Over 17% more minority pupils pass the BSF if the pupils are in S rather than in R (or RA).*

The statistical significance question seems to be resolved in class size issues. There remains the "educational" significance question. Often "educational" significance is dealt with by reviewing the "effect sizes." Effect size is one way to see *how much* the gain is relative to a standard deviation. With the CRT, an educational effect might be the percent passing, as percent has a standard of 100. Effect sizes favoring S in STAR range from .08 (in K) to .40 (in grade 3) for minority pupils. Generally, the positive STAR effect sizes for pupils in S are in the .20 to .27 range. (See Table 5.)

---

*The STAR Consortium used an external advisory board and an external consultant to conduct independent analyses of STAR data. Project and external analyses were confirmatory. The achievement analysis involved Stanford Achievement Tests, or SAT, and Tennessee's criterion-referenced BSF tests. The Consortium chose SESAT II over SESAT I since Tennessee (K) objectives correlated better with SESAT II than with SESAT I, and SESAT II offered a higher "ceiling," allowing pupils to show greater gain. The Consortium also chose "comparison" schools selected from STAR districts which already used the SESAT II, SAT, and other tests. Analyses of STAR results with comparison-school results are in process (1994).

Table 3
Analysis of Variance for Cognitive Outcomes, STAR, Grades K-3.
Sig. Levels $p < .05$ or Greater are Tabled.

Effect/[a] Grade		Reading			Mathematics		
		Multi-variate[b]	SAT[c] Read	BASF Read	Multi-variate[b]	SAT Math	BASF Math
Location (L)	K		.02			.05	
	1	.01	.06		.05		
	2	.001	.001	.001		.001	.001
	3	.001	.001	.001	.001	.001	.001
Race (R)	1	.001	.001	.001	.001	.001	.001
	2	.001	.001	.001	.001	.001	.001
Type (T)	K		.001			.02	
	1	.001	.001	.001	.001	.001	.05
	2	.001	.001	.05	.001	.001	.05
	3	.001	.001	.001	.001	.001	.001
SES	K		.001			.02	
Loc X Race	1	.05		.05			
Loc X Type	K-3	All N/S. The class-size effect is found equally in all locations--Inner City, Suburban, Urban, and Rural schools. (Tabled as important.)					
Race X Type	1	.05	.05	.01			
LxRxT	1			.05			.01
LxTRxT	2	.05	.01	.05	.05	.05	.01

NOTE: [a]The nonorthogonal design required tests in several orders (Finn & Bock, 1985). Results were obtained as follows: Each main effect was tested eliminating both other main effects; loc x race tested eliminating main effects and loc x type tested eliminating main effects and loc x race; race x type tested eliminating main effects and other two-way interactions, and loc x race x type tested eliminating all else (Finn & Achilles, 1990). [b]Obtained from F-approximation from Wilks' likelihood ratio. Essentially, no statistically significant differences were obtained on the self-concept and/or motivation (SCAMIN) measures. No training main effect, or training-by-type interaction. Trained and untrained teachers did equally well across all class types and the (S) advantage (and absence of Aide effect) is found equally in all four locations for trained and untrained teachers. (S) advantage and all effects for total class generally apply equally to white and nonwhite pupils, especially in grade 2. The race difference was statistically significant for all measures and multivariate sets, but *not* for most interactions (LxR, TRxR, TxR, LxT, R, or TRxTxR). (S) Significantly better than (R, RA) on all tests; no R vs RA tests significant. This basic data table appears in other articles and conference reports by the same authors.

Table 4
Average Percent of Pupils passing
BSF Reading: Grade 1, STAR

Status	Grade	Class Type Small	Class Type Reg.	Difference (S-R) or (S) Advantage
Minority	1	65.4%	48.0%	17.4
Non-Minority	1	69.5%	62.3%	7.2
Difference		4.1%	14.3%	

Table 5
Estimates of (S) Effect Sizes, Using (S) and
(R & RA) ÷ 2* for White (W), Minority (M),
and All Pupils, K, 1, 2, and 3, STAR, 1985-1989

Scale	Group	Grade K	1	2	3**
SAT TESTS					
Total	W	-	.17	.13	.17
Read	M	-	.37	.33	.40
	All	.18	.24	.23	.26
Total	W	.17	.22	.12	.16
Math	M	.08	.31	.35	.30
	All	.15	.27	.20	.23
BSF Tests					
BSF	W	-	4.8%	1.6%	4.0%
Read	M	-	17.3%	12.7%	9.3%
	All	-	9.6%	6.9%	7.2%
BSF	W	-	3.1%	1.2%	4.4%
	M	-	7.0%	9.9%	8.3%
	All	-	5.9%	4.7%	6.7%

* Effect size is difference divided by the appropriate standard deviation (for groups or totals). The BSF percents are calculated from differences of groups in percent passing. No BSF tests were given in K. Grade 2 computed on untrained teachers only (N = 273).

** Grade 3 was computed on Total Language Test results.

Phase II. The Lasting Benefits Study (LBS)

What happens when STAR pupils who benefitted from S in K-3 return in grades 4 and later to "regular" classes? Weikart (1989) and material in *Futurist Magazine* ("Education," 1990) point out the lasting benefits of early intervention. The STAR database provides the opportunity for a longitudinal study of benefits of early small-class involvement. The LBS is primarily a process to follow pupils who were in STAR in the S, R, RA conditions. Analyses use pupil test scores and behavioral indicators of school efforts. The fourth-grade analysis included 4,230 pupils. (They were identified by class type in at least grade 3.) Of those, 1,312 were S, 1,250 were R, and 1,568 were RA. Fifth-grade analyses included 4,649 pupils: 1,578 (S), 1,467 (R), and 1,604 (RA). The LBS lacks the design strengths of STAR; LBS is "field research" while STAR was a true "experiment." Nevertheless, the LBS results are informative and an important contribution to the analysis of class-size intervention and public policy decision making.

Scaled-score means for STAR class types (S, R, RA) were compared through multivariate analysis of variance (MANOVA) for unequal n's using the MULTIVARIANCE program (Finn & Bock, 1985). The analysis examined mean differences among three class types, the mean differences among four school locations (rural, urban, suburban, inner city), and the interaction between class types and locations. Using the basic STAR analysis design, three achievement subsets for the LBS were compared separately. Two subsets include scores from both the NRT and CRT components of the Tennessee Comprehensive Assessment Program or TCAP. Set 1 included Total Reading (NRT scores), Total Language (NRT scores), and the number of domains mastered in Language Arts (CRT). Set 2 consisted of Total Math (NRT scores), Total Science (NRT scores), and the number of domains mastered in Mathematics (CRT). Set 3 included Study Skills (NRT) and Social Science (NRT) scores. (See also Finn et al., 1989/1992). By grade 5 some pupils entered middle schools and the analysis by location no longer seemed feasible.

The LBS analysis yielded clear and consistent results. Students previously in a small-size STAR class demonstrated in every location that they had statistically significant ($p \leq .01$) advantages over R and RA pupils on every set of measurements. The greatest achievement advantages (grade 4) were for inner-city and suburban classes (Table 6). For grades 4 and 5, all S v. R contrasts were significant ($p \leq .01$); no R v. RA contrast was significant.

Table 6
LBS Results, Grade 4 (1989-90) and Grade 5 (1890-91) on TCAP.
Summary of Class Effects Analysis Using Mean Scores of Sets

	Set 1 Verbal		Set 2 Math/Sci		Set 3 SocSci/Study	
	4	5	4	5	4	5
Loc (urban, etc.)	$p \leq .001$	N/A	$p \leq .001$	N/A	$p \leq .001$	N/A
Type (S, R, RA)	$p \leq .001$	$p \leq .01$	$p \leq .001$	$p \leq .01$	$p \leq .001$	$p \leq .01$
Loc X Type	NS	N/A	NS	N/A	NS	N/A

(Results found in all locations equally)

Loc. differences on all sets favoring S in the location, but major difference is due mostly to lower-performing inner-city pupils. Type differences favor S. R vs RA contrasts NS. Loc X Type class-type differences are the same in all locations. (Nye et al., 1991, 1992).

The Project STAR results indicated substantial educational benefits for students in small classes. *The positive effects from involvement in a small-size class still remain pervasive two full years after students returned to regular-size classes.* The LBS students who had attended small STAR classes had an educationally and statistically significant advantage over LBS students who had attended R or RA STAR classes. This advantage can be measured by the TCAP scaled-score differences between S and R classes, and between the RA and R classes as shown in Table 7. *Students from the S classes retained their academic advantage.*

Table 8 provides estimates of the S and RA class effect sizes, grades 4 and 5, 1989-90 and 1990-91. Effect sizes ranged from .11 to .34 for the S/R contrast. The R/RA contrast shows effect sizes ranging from -.02 to -.09 (Finn et al., 1989/1992; Nye et al., 1991, 1992). The significant advantages for LBS fourth- and fifth-grade students who had been in STAR small classes form a strong pattern of consistency. Small-class students outperformed R and RA class students on every achievement measure.

As part of the LBS analysis Finn et al. (1989/1992) reported differences in student *participation* based on prior class-size experiences (S, R, RA). [Details of the participation idea appear in Finn (1989) and in Finn and Cox (1992).] Essentially, according to Finn (1989), increased student participation in school reflects a decreasing tendency for student alienation and dropout in later years. Opportunities for student participation (e.g., clubs, service projects, music, athletics) can be established and operated by those in schools--teachers and administrators. Participation also includes the pupil's active involvement in classroom activity.

Table 7
LBS: Grades 4 and 5. TCAP, Scaled Score Differences and the Differences in Mean Number of Domains Mastered between S and R Class Students and between RA and R Class Students. Means are tabled in Appendix B of the Technical Report (Nye et al., 1991, 1992).

Measures	1989-90 (4th)		1990-91 (5th)	
NRT	S vs R	R vs RA	S vs R	R vs RA
Total Reading	5.61	-2.23	10.53	.10
Total Language	4.99	-.73	8.21	-1.03
Total Math	4.87	-2.29	8.08	-.34
Science	5.69	-1.47	8.99	-2.66
Social Sciences	6.13	-.195	8.14	-1.31
Study Skills	10.10	-2.15	10.62	-.85
CRT (Domains Mastered)				
Language Arts	.25	-.18	.84	.07
Mathematics	.35	-.09	.68	.16

Table 8
LBS: Grades 4 and 5, 1989-90; 90-91.
TCAP. Estimates of S and RA Effect Sizes

Measures	1989-90 (4th)		1990-91 (5th)	
NRT	S v R	R v RA	S v R	R v RA
Total Reading	.13	-.05	.22	.00
Total Language	.13	-.02	.18	-.02
Total Math	.12	-.06	.18	-.01
Science	.12	-.03	.17	-.05
Social Science	.11	-.04	.17	-.03
Study Skills	.14	-.03	.18	-.01
CRT				
Language Arts	.11	-.09	.34	.03
Mathematics	.16	-.04	.28	.07

Finn et al. (1989) assessed a grade 4 subset of STAR pupils by asking their teachers to rate them on the 25-item Pupil Participation Questionnaire on a 5-point range from (1) "never" to (5) "always." Teachers rated pupils on three behavioral scales (Finn et al., 1989/1992).

> Nonparticipatory Behavior (e.g., "Annoys or interferes with peers' work"), Minimally Adequate Effort (e.g., "Pays attention in class"), and Initiative Taking (e.g., "Does more than just the assigned work"). (p. 78)

Teachers rated pupils in their classes who had participated in one of three STAR conditions for three years (grades 1-3). The 258 teachers in 74 schools rated 2,207 pupils. Using the STAR and LBS MANOVA design, scores on the three participation scales--Effort, Initiative, and Nonparticipatory Behavior--were simultaneous criterion variables (p. 79).

> [Location ($p \leq .05$); Class type ($p \leq .0001$); Loc x Type ($p \leq .05$)] (p. 79).

According to Finn et al. (1989/1992):

> The particular contrast of small-class with regular-class students was statistically significant at $p \leq .05$ using a multivariate test and at p-values of .05 or .01 on individual scales. Pupils who had attended small classes were rated as having superior modes of participation in grade 4 in comparison to their peers. (p. 81)

The participation effect sizes (.11 to .14) were similar to effect sizes found in LBS achievement analyses (.11 to .16). The R/RA contrast was not significant. To date, the LBS study shows that the STAR small-class benefit is retained consistently *two full years after STAR ended.* There is also the added benefit of increased participation behavior--positive behavior linked to staying in school (Finn, 1989). This LBS analysis links the desired participation behavior to higher academic achievement on measures used in LBS. (Although not obtained for the grade 5 analyses, LBS researchers plan to assess participation again.)

Building upon the database provided by STAR, LBS is showing that *early* small-class involvement (e.g., 1:15) has continuing benefits (note also Weikart, 1989). This does, in effect, deflect some criticism of the *cost* of reduced class size, since the benefits are spread out over more years than simply during the years of the class-size reduction.

Phase III. Project Challenge as Policy Implementation

To help pupils in some of Tennessee's poorer counties, the state provided funding and incentives for local district leaders to use various strategies to improve pupil performance. Beginning in 1989, one option--called Project Challenge--was to reduce the class size in 17 districts in grades K-3 to approximately 1:15. Project Challenge put into practice results of the statewide STAR experiment.

Prior to the 1989-90 school year, Tennessee pupils generally took the Stanford Achievement Tests (SAT) as the state testing format. Beginning in 1989-90 students in selected grades began taking the Tennessee Comprehensive Assessment Program, or TCAP. The TCAP includes both NRT and CRT components. Since no special testing was done for Challenge, extant data and regular testing processes were used in the evaluation plan. Test data and results for all discussions are for grade 2, the first grade level for regular TCAP testing on a statewide basis.

The Tennessee SEA needed some idea if the class size reduction (1:15) *seemed* to be helping student achievement in the 17 counties. Since in Challenge there was no "experiment" with random selection or assignment, no special testing, etc., an evaluation is essentially an after-the-fact (post hoc) review and analysis of grouped (e.g., school system) data, using the available second grade test results. There is no sure way to attribute any gain (or loss) to Challenge (e.g., class-size reduction) if other special "interventions" were taking place at the same time in the same grades.

There may be other systematic threats to validity, too. Grouped data by grade level are subject to any variation in student ability by classes or grades. Gains or losses in one year may be the result of very good (or very poor) student ability, excellent teaching, test variation, etc. Only with several years of results can a trend become evident. Experience with STAR and LBS can help in Challenge.

Thus, since testing changed in 1989-90 and Challenge began in 1989-90, use of 1989-90 second grade TCAP results as the baseline data for Challenge means that the second graders in 1989-90 already had one year of CHALLENGE (that is, 1989-90 data are baseline *after* one year of treatment). Use of 1990 TCAP as "baseline" even when pupils had one year of "treatment" seemed preferable to using the pre-Challenge but not comparable SAT results for second graders. The 1989-90 data reflect one year (only grade 2) of time in Challenge for the pupils. The 1990-91 data reflect those pupils who had Challenge class-size reduction (1:15) in grades 1 (1989-90) *and* 2 (1990-91), etc. (See Table 9.)

Table 9

Summary Table of Students in Project Challenge (TN: 1990-93) and Years of Testing Using TCAP Tests to Analyze Challenge Successes*

Grade 2 pupils' experience in Challenge (in years) by grades at time of testing

Test Date	Years in Challenge	Grades of Challenge	Test Used/Grade
1990	1	Grade 2 only	TCAP, Grade 2
1991	2	Grades 1 and 2	TCAP, Grade 2
1992	3	Grades K-2	TCAP, Grade 2
1993, etc.	3	Grades K-2	TCAP, Grade 2

* Challenge reduces class size (1:15) in grades K-3.

Although there clearly are limitations, one fairly simple way to see if Challenge systems as a group ($n = 17$) seemed to be benefitting from the treatment (i.e., 1:15) is to consider the rankings (or the aggregate rankings) of the 17 Challenge systems among all Tennessee systems ($n = 138$). This was done for reading and for math by adding the rankings of the 17 systems (using data provided by the SEA), then dividing by 17 to get the "average" ranking in 1989-90 (baseline) and then in subsequent years (e.g., 1990-92). Since a rank of "one" is best, a gain is achieved when the aggregate (and average) ranks become *lower*. With a total of 138 systems, the state average rank would be 69.

Data in Table 10 show that, on average, the Challenge systems *moved up* 5.3 ranks in reading and 6.6 ranks in math from 1989-90 to 1990-91. The average Challenge system (1990-91) was at 94 in reading and 79 in math, still below the state average (69). However, a different picture emerges in the 1991-92 data when the Challenge pupils had three years of small class treatment beginning in K (the year they started school). Note that in math the average Challenge system is now *above* the state average rank and that reading continues to rise.

Table 10

Rankings of Challenge Districts (n = 17) of 138 TN School Systems Based on Grade 2 TCAP Scores (Reading and Math). Average rank is 69.

	Reading			Mathematics		
	89-90	90-91	91-92	89-90	90-91	91-92
Sum of Ranks	1681	1591	1477	1448	1336	1011
÷ by 17	98.9	93.6	86.9	85.2	78.6	59.5*
Difference		(+90)	(+114)		(+112)	(+325)
÷ by 17		5.3 RK	6.7 RK		6.6 RK	19.1 RK

*Above state average.

A second procedure is to convert the district average scores to *z*-scores and then to consider how the 17 Challenge systems' grade 2 average scores in reading and math deviate (e.g., in terms of standard deviation units) from the state average. Although the average *z*-scores for reading for 1990, 1991, and 1992 TCAP results are below the state average, the .23 and .13 standard deviation gains moved these 17 systems closer to the state mean from 1990 to 1992. The *z*-score gains in math (.26 and .38) from -.34 to +.30 show that the average math rank for Challenge is above the state norm. (See Table 11.)

Gains in rankings and in *z*-score comparisons show that, on average, the second grade TCAP results are going in the desired direction; student scores are getting better as the systems move up relative to the state average. Subsequent analyses will see if the trend continues.

Table 11
Comparison of Challenge Systems (n = 17) Average z-Scores for Reading and Math, Grade 2, TCAP Results

	Reading			Mathematics		
	89-90	90-91	91-92	89-90	90-91	91-92
z-Score Difference	-.75	-.52 Gain (.23)	-.39 Gain (.13)	-.34	-.08 Gain (.26)	.30 Gain (.38)

Discussion

Class size reduction, as a treatment or intervention, is really a one-time event. That is, the treatment is when the student first experiences the reduction from regular (e.g., 1:28) to small (1:15); the ensuing years are a *continuation*, but not a separate treatment.

Challenge systems gained in the state rankings, but the magnitude of the gains was less than the demonstrated gains in STAR until the analyses included pupils who *started* school in K in 1:15. Although consistent in all STAR conditions (S, R, RA), pupil assignment in STAR (random) was different from regular pupil assignment practices. Did pupil *random assignment* positively influence STAR results in all or in some STAR conditions? Additional analyses of the STAR database are helping unravel this interesting question.

The LBS results show the continuing benefits of a pupil's participation in the small class. Post hoc analyses of important elements of schooling other than achievement (e.g., participation) suggest a small-class influence here, too. Continuing analyses through LBS will add to information provided by other longitudinal studies (e.g., Weikart, 1989) of important social benefits of early primary and pre-primary interventions. Zigler (1992) emphasizes that *in spite* of continual strong evidence of success of Head Start, the funding continues to erode and "$250 million . . . was dropped from the emergency aid bill" (p. 15). Children clearly are less important than other budget items! [In an attempt to deal with California's budget crisis (7/92) Governor Wilson suggested eliminating kindergarten, at least for one year.]

Since LBS shows continuing benefits in pupil achievement *after* small-class involvement, will small-class involvement for one or two years (rather than STAR's four years) provide a sound base to help pupils get started well in school? If so, STAR results were strongest in K and 1, suggesting that these should, at a minimum, be the years of the small-class intervention. The early primary heterogeneous classes provided by the STAR random assignment and STAR's seeming ability to help minority pupils close the achievement gap are promising areas for LBS analyses. The Ramey (1992) model may help here.

Although STAR's greatest gains were in K-1 and the gain was not as large in grades 2-3, the initial gain is maintained and enhanced through third grade. Thus, while K-1 students really benefit from small classes, students in grades 2-3 continue to benefit (or, if they encounter small classes for the first time in grades 2-3, get initial benefits) from small classes. Small classes allow for more developmentally appropriate curriculum, instruction, and parent involvement. Small classes are especially important for children through third grade and for teachers who increasingly must deal with greater pupil disadvantagement and diversity in single grades.

Results of STAR (the experiment) provide *clear* evidence of ways to improve schooling in early primary grades. Given the added needs of children entering schools in the 1990s (e.g., Hamburg, 1992; Hodgkinson, 1991), the use of small classes may become imperative for later school success. We have found a *way* to improve schooling; do we have the *will*? The STAR experiment results have held up in field research and policy conditions (e.g., LBS, Challenge) and are continuing to show added, continuous benefits. With this much evidence, leaders in Tennessee and in other states are implementing class size reductions. How much more evidence do other policy makers need before they apply sound research results to school improvement?

Results of research covering 1985-1994 describe one effective way to improve education. Should these and similar studies be seen simply as studies in class size reduction? Perhaps they are better cast as trying to find the right class sizes to help solve Bloom's (1984) "two-sigma" problem--trying to match the size of the instructional unit to the job to be done. The results suggest ways to move from assembly-line, industrial-age schooling to caseload, information-age learning activities. Small is definitely far better in the long run.

References

Bloom, B. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership, 41*(8), 4-17.

Cahen, L. S., Filby, N., McCutcheon, G., & Kyle, D. W. (1983). *Class size and instruction.* New York: Longman.

Education: Long-term benefits of preschool. (1990). *The Futurist Magazine, 24*(2), 49.

Education Research Service or ERS. (1978). *Class size: A summary of research.* Arlington, VA: Author.

Education Research Service or ERS. (1980). *Class size research: A critique of recent meta-analysis.* Arlington, VA: Author.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59*(5), 117-142.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*(3), 557-577.

Finn, J. D., Achilles, C. M., Bain, H. P., Folger, J., Johnston, J., Lintz, M. N., & Word, E. (1990). Three years in a small class. *Teaching and Teacher Education, 6*(2), 127-136.

Finn, J. D., & Bock, R. D. (1985). *MULTIVARIANCE VII user's guide.* Mooresville, IN: Scientific Software, Inc.

Finn, J. D., & Cox, D. (1992). Participation and withdrawal among fourth-grade pupils. *American Educational Research Journal, 29*(1), 141-162.

Finn, J., Zaharias, J., Fulton, D., & Nye, B. (1989). Carry-over effects of small classes. *Peabody Journal of Education, 67*(1), 141-162. (Published in 1992).

Folger, J., (Ed.). (1989). Project STAR and class size policy. *Peabody Journal of Education, 67*(1). (Published in 1992).

Glass, G. V., & Smith, M. L. (1978). *Meta-analysis of research on the relationship of class size and achievement.* San Francisco: Far West Laboratory for Educational Research and Development.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size, research and policy.* Beverly Hills: Sage Publications.

Hamburg, D. A. (1992). *Today's children.* New York: Time Books, Random House.

Hodgkinson, H. (1991). Reform vs. reality. *Phi Delta Kappan, 73*(1), 8-16.

Mitchell, D. E., Beach, S. A., & Badarak, G. (1989). Modeling the relationship between achievement and class size: A re-analysis of the Tennessee Project STAR data. *Peabody Journal of Education, 67*(1), 34-74. (Published in 1992).

Mitchell, D. E., Carson, C., & Badarak, G. (1989). *How changing class size affects classrooms and students.* University of California, Riverside: California Educational Research Cooperative.

Mueller, D. J., Chase, C. I., & Walden, J. O. (1988). Effects of reduced class sizes in primary classes. *Educational Leadership, 45*, 48-50.

Nye, B., Zaharias, J., Fulton, D., Achilles, C. M., & Hooper, R. (1991, 1992). *The lasting benefits study; Technical reports (grades 4 and 5).* Nashville, TN: Tennessee State University Center for Excellence.

Orlich, D. C. (1991). Brown v. Board of Education: Time for a reassessment. *Phi Delta Kappan, 72*(8), 631-632.

Ramey, M. (1992). *Classroom characteristics related to ethnic achievement gap reduction.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Robinson, G. E. (1990). Synthesis of research on the effects of class size. *Educational Leadership, 47*(7), 80-90.

Robinson, G. E., & Wittebols, J. H. (1986). *Class size research: A related cluster analysis for decision making.* Arlington, VA: Educational Research Service, Inc.

Shapson, S. M., Wright, E. N., Eason, G., & Fitzgerald, J. (1980). An experimental study of the effects of class size. *American Educational Research Journal, 17*, 141-152.

Slavin, R. E. (1989). Achievement effects of substantial reductions in class size. In R. E. Slavin (Ed.), *School and classroom organization* (pp. 247-257). Hillsdale, NJ: Lawrence Erlbaum Associates.

Slavin, R. E. (1990). Class size and student achievement: Is small better? *Contemporary Education, LXII*(1), 6-12.

Tomlinson, T. M. (1988). *Class size and public policy: Politics and panaceas.* Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Tomlinson, T. M. (1990). Class size and public policy: The plot thickens. *Contemporary Education, LXII*(1), 17-23.

Weikart, D. P. (1989). *Quality preschool programs: A long-term social investment.* Occasional Paper Number 5. Ford Foundation Project on Social Welfare and the American Future. New York: The Ford Foundation. (28 pp.)

Word, E., Johnston, J., Bain, H., Fulton, B., Zaharias, J., Lintz, N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/teacher achievement ratio (STAR): Tennessee's K-3 class size study.* Final report and final report summary. Nashville, TN: Tennessee State Department of Education.

Zigler, E. (1992, June 27). Head Start falls behind. *New York Times*, p. 15.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 21-28

Aspirations of Minority High School Seniors in Relation to Health Professions Career Objectives

William A. Thomson, Leslie Michel Miller, Bernice Ochoa Shargey, James P. Denk, and Bruce Thompson

Efforts to attract minority students to health professions careers are important. Minority groups have greater needs for health care and tend to be less willing to seek needed care, partly in reaction to underrepresentation of minority groups within the health professions. The present study explored career-related perceptions of the minority seniors at two High Schools for the Health Professions located in disparate areas of Texas.

Schools can impact the career choices made by high school students in at least two ways: through curriculum revisions and innovations such as magnet schools, and, perhaps more directly, through career counseling. Modern career counseling reflects the paradigm shift (Super, 1951) recognizing that counselors play an important active role in helping clients to make choices that lead to satisfaction both for self and for society. One important area where impacts on career choices are urgently needed involves the potential for high school counselors and curricula to work together to help redress the serious and avoidable underrepresentation of minorities within the health professions (Health Resources and Services Administration, 1984; Mingle, 1987).

Data on ethnic minority representation in the health professions reflect striking disparities between the percentages of African-American and Hispanic persons in almost all health professions and their representation in the general population (Committee on Allied Health Education and Accreditation, 1991; Health Resources and Services Administration, 1990). For example, while the 1990 U.S. census indicated that the national population was 11.8% African-American and 9.0% Hispanic, African-American (6.6%) and Hispanic (2.7%) citizens together accounted for only 9.3% of U.S. medical school matriculants in 1991 (Association of American Medical Colleges, 1992). Furthermore, in 1985-1986, enrollment in the nation's registered nursing programs was 10.3% African-American and 2.7% Hispanic, while in 1988-1989, first year enrollment in dentistry was 6.9% African-American and 7.6% Hispanic (Health Resources and Services Administration, 1990).

And similar, if not even more severe, disparities exist in the allied health professions (Institute of Medicine, 1989). Even the high-demand professions of physical therapy and occupational therapy included only 2.1% and 3.3% African-American and 0.9% and 1.1% Hispanic citizens, respectively. In fact, only in laboratory technician (11.1% of overall practicing professionals) and respiratory therapy (10.0%) did African-American representation approach the percentage of African-Americans in the population. Finally, Hispanic representation in allied health fields remains far below the percentage of Hispanics in the general population. At 4.9%, Hispanic representation has been highest in respiratory therapy.

This profile is disturbing, because people are less likely to seek health care when their ethnic groups are under-represented among health care providers (U.S. Department of Health and Human Services, 1985). With respect to health care for Hispanics, for example, Garcia and Ramon (1988) argued that:

> The underrepresentation of Hispanics in the health-care professions carries with it both a human and political toll. The intent of parity is founded in the notion of equality. However, a motivating force in the drive to reach parity is

---

William A. Thomson is Associate Professor of Community Medicine and Head of the Division of School-Based Programs at Baylor College of Medicine. Leslie Michel Miller is Research Assistant Professor of Community Medicine at Baylor College of Medicine and Faculty Fellow of Education at Rice University. Bernice Ochoa-Shargey is Assistant Professor of Community Medicine at Baylor College of Medicine and Dean of Instruction at the Houston High School for Health Professions. James P. Denk is Editor and Administrative Associate in the Department of Community Medicine at Baylor College of Medicine. Bruce Thompson is Professor of Education and Distinguished Research Fellow at Texas A&M University, and Adjunct Professor of Community Medicine at Baylor College of Medicine. Correspondence and requests for reprints should be addressed to Dr. William A. Thomson, Baylor College of Medicine, Suite 545, 1709 Dryden, Houston, TX 77030-3498.

> the concept of service--more Hispanic health-care professionals will improve health-care services received by the Mexican-American community of Texas. (p. 242)

Because these views generalize to other ethnic minority groups and to other geographic areas, the Institute of Medicine's (1989) Committee to Study the Role of Allied Health Personnel recommended that:

> The recruitment of minority students is a particular concern for several reasons: minorities represent a relatively untapped source of human power; their representation in the population as a whole is increasing; and minority professionals are more likely to serve underserved populations.
>
> There have been a number of attempts to recruit and retain minorities in the health professions. The lessons from successful models suggest that interventions must occur early in a student's life and continue through the academic career. (p. 8)

A recent report from the Pew Health Commission (Shugars, O'Neil & Bader, 1991), *Healthy America: Practitioners for 2005*, supports this view, adding that

> minorities, previously underrepresented in the health professions, will become a large part of the pool of potential applicants to health professional schools. Health professions in general and health professional educators in particular will need to understand and relate to the special needs of this growing segment of society. (p. 7)

So, too, will career counselors.

The purpose of the present study was to explore the perceptions of minority, high-school, senior students as regards career choices involving the health professions. Understanding such perceptions may be useful to career counselors working with clients, such as the participants in the present study, and may offer some guidance for curricular change.

Specifically, the present study was conducted to address three research questions. First, what, if any, ethnic-group differences are there as regards the career-related perceptions and choices of minority students? Second, what factors, if any, predict the health-career choices of minority students as high school seniors? Third, what factors, if any, predict the decisions of minority students to change career objectives while they are enrolled in health-professions magnet schools?

Method

Subjects

The participants all were enrolled in one of two magnet, alternative high schools for health professions careers. One high school is located in an urban area--Houston, Texas; the second high school for health professions is located in the Rio Grande "Valley" of Texas, an area near the Texas-Mexico border, and which has a disproportionately higher Hispanic populace than either the urban school district or the population of the country as a whole. As magnet schools, both programs draw students from a broad geographic area, i.e., they do not limit enrollment only to persons living in the neighborhood nearby the school building. Both schools also consciously strive to maintain ethnically diverse student censuses.

The features of these high schools for the health professions have been described elsewhere (e.g., Butler, Thomson, Morrissey, Miller & Smith, 1991; Miller, LaVois & Thomson, 1991; Thomson, Holcomb & Miller, 1987; Thomson, Smith, Miller & Shargey, 1991). The most relevant aspect of the schools, as regards the present study, is that students initially voluntarily enter the schools because they wish to explore the nature of careers in the health professions and/or because they wish to acquire the high school preparation requisite for such career choices.

Although *all* the seniors enrolled in the Houston High School for the Health Professions and the South Texas High School for the Health Professions participated in the study, relatively few nonminority students (12 in both schools) were represented in the study. Given the disproportionately small representation of these nonminority students, and the emphasis in the present study on dynamics involving the minority senior high school students, the decision was taken to exclude nonminority students from the analyses reported here. Table 1 profiles the two samples.

Instrumentation

The instrument employed in the study was a derivative of the measure employed in a recent, national study in a series of studies of college freshmen (Astin, Dey, Korn & Riggs, 1991). Thus, the instrument has been thoroughly investigated and refined. The instrument asked about (a) current career goals; (b) career choice changes, if any; (c) influences on career decisions; (d) educational goals; (e) career choice satisfaction; and (f) perceived obstacles to career objectives. The items used in the current analysis are summarized within the tables of this report. A copy of the complete instrument is available from the senior author.

Table 1
Characteristics of the Sample

Characteristic	Urban TX	South TX	Total
Gender			
Female	97 (70.8%)	41 (63.1%)	138 (68.3%)
Male	40 (29.2%)	24 (36.9%)	64 (31.7%)
Total	137	65	202
Ethnicity			
African-American	79 (57.7%)	0 (0.0%)	79 (39.1%)
Asian	25 (18.2%)	0 (0.0%)	25 (12.4%)
Hispanic	33 (24.1%)	65 (100.0%)	98 (48.5%)
Total	137	65	202
Career Goal			
Allied Health	16 (11.7%)	9 (13.8%)	25 (12.4%)
Business	13 (9.5%)	2 (3.1%)	15 (7.4%)
Dentistry	7 (5.1%)	2 (3.1%)	9 (4.5%)
Medicine	64 (46.7%)	15 (23.1%)	79 (39.1%)
Nursing	14 (10.2%)	21 (32.3%)	35 (17.1%)
Vet Medicine	2 (1.5%)	1 (1.5%)	3 (1.5%)
Other, not health	18 (13.1%)	11 (16.9%)	29 (14.4%)
Undecided	3 (2.2%)	4 (6.2%)	7 (3.5%)
Total	137	65	202

Results

The research questions in the present study involved differences across categorical groupings. The analytic method used was multivariate, so as to avoid inflation of experimentwise error rate and to represent the full network of relationships among variables (Thompson, 1992). Discriminant function analysis (Huberty & Wisenbaker, 1992) was employed to address these research questions. Since all the discriminant function analyses in the present study involved a single grouping variable, the discriminant results are equivalent to one-way MANOVAs, but provide more descriptive information, useful in formulating interpretations, than does MANOVA.

Prior to conducting analyses addressing the study's three research questions, a preliminary ancillary analysis was conducted to explore differences between the seniors enrolled in the two high schools. These analyses compared the two groups as regards 10 factors influencing their decisions to seek additional education, as well as perceptions of concerns regarding financing future education, confidence about ability to achieve current career goals, and satisfaction with current career choices. The two groups did not differ to a statistically significant degree ($\lambda = .94$, $\chi^2 = 11.14$, $df = 13$, $p = .60$) as regards these 13 variables. All further analyses reported here were conducted by pooling participants across school sites.

A second ancillary analysis explored possible gender differences within the sample as regards mean differences on these same 13 variables. Again, these two groups did not differ to a statistically significant degree ($\lambda = .92$, $\chi^2 = 15.30$, $df = 13$, $p = .29$).

The study's first research question involved exploring group differences across the three minority groups represented in the sample. The same 13 variables were employed in this analysis. Since there were three groups in this analysis, two (3 - 1) discriminant functions were computed (Huberty & Wisenbaker, 1992). However, only the first lambda value was statistically significant ($\lambda = .79$, $\chi^2 = 43.10$, $df = 26$, $p = .02$).

Table 2 presents the standardized function coefficients and the structure coefficients associated with Function I in this analysis. Standardized function coefficients are directly analogous to the beta weights in regression analysis. Structure coefficients are correlation coefficients between scores on the interval variables and scores on discriminant functions, calculated using the function coefficients as weights for the interval variables. Both sets of coefficients are important in interpreting regression results, and are also important in interpreting discriminant analysis results (Thompson & Borrello, 1985). The Table 2 entries are presented in descending order of the absolute value of the variables' structure coefficients.

The Function I centroids (i.e., mean discriminant function scores) for the African-American, the Asian, and the Hispanic students were B0.48, +0.60, and +0.20, respectively. These results indicate that the African-American students were most different from the Asian students, as regards the first discriminant function.

Table 2
Discriminant Coefficients Involving Ethnicity Differences

Variable	β	r_s
To be able to make more money among reasons most important in deciding to further education	-.83	-.58
Feel will be able to achieve present career goal	.42	.37
To achieve my career objectives among reasons most important in deciding to further education	-.14	-.28
To become a more cultured person among reasons most important in deciding to further education	.53	.25
To get away from home among reasons most important in deciding to further education	-.22	-.25
To get a better job among reasons most important in deciding to further education	-.10	-.23
Parents' wishes among reasons most important in deciding to further education	.28	.21
Level of concern about financing further education	.03	.14
To prepare for graduate school among reasons most important in deciding to further education	-.10	-.14
Level of satisfaction felt since identifying current career goal	.02	.10
To learn more about things among reasons most important in deciding to further education	.28	.08
To improve reading and study skills among reasons most important in deciding to further education	-.11	-.08
To gain general education among reasons most important in deciding to further education	.04	-.05

Note. "**β**" = standardized discriminant function coefficients, directly analogous to regression beta weights. Entries are presented in descending order of the absolute value of the variables' structure coefficients.

The study's second research question involved predicting the career choices of the 195 seniors who were able to articulate a career goal. This variable involves dynamics of change in career objectives, since the students had some commitment to or at least interest in the health professions when they voluntarily matriculated to these specialized high schools. For the purposes of this analysis the career goals were divided into three categories: (a) medicine ($n_1 = 79$); (b) health, but not medicine ($n_2 = 72$); or (c) business or some other non-health career ($n_3 = 44$).

The first lambda value was statistically significant ($\lambda = .76$, $\chi^2 = 50.00$, $df = 26$, $p < .01$), and so was the second lambda value ($\lambda = .88$, $\chi^2 = 23.82$, $df = 12$, $p = .02$). Table 3 presents the standardized function coefficients and the structure coefficients associated with both discriminant functions in this analysis.

The Function I centroids (i.e., mean discriminant function scores) for the medicine, the non-medicine health, and the non-health choices on Function I were +0.46, -0.27, and -0.38, respectively. These results indicate that the students with medicine as an objective differed most from both the other groups, although somewhat more with respect to the non-health group.

The Function II centroids for the medicine, the non-medicine health, and the non-health choices on Function II were -0.05, +0.40, and -0.57, respectively. The fact that this was the second function indicates that group differences associated with this function were smaller in magnitude than those associated with Function I. The Function II centroids indicate that this function is most useful in explaining differences between the non-medicine health group as against the non-health group, ignoring the students with medicine as a career objective.

Table 3
Discriminant Coefficients for Two Discriminant Functions When Predicting Career Goals

Variable	Function I		Function II	
	β	r_S	β	r_S
To get a better job among reasons most important in deciding to further education	-.60	-.45	.40	.22
To get away from home among reasons most important in deciding to further education	-.44	-.44	.16	.02
Feel will be able to achieve present career goal	.63	.41	-.14	.03
To be able to make more money among reasons most important in deciding to further education	-.26	-.32	-.45	-.20
To prepare for graduate school among reasons most important in deciding to further education	.49	.27	-.34	-.20
To improve reading and study skills among reasons most important in deciding to further education	.16	.14	.51	.41
To become a more cultured person among reasons most important in deciding to further education	-.11	-.12	-.53	-.35
Level of concern about financing further education	-.05	.09	.56	.31
To gain general education among reasons most important in deciding to further education	.27	.14	.12	.27
To learn more about things among reasons most important in deciding to further education	.09	.09	.43	.22
To achieve my career objectives among reasons most important in deciding to further education	.08	.02	.20	.16
Level of satisfaction felt since identifying current career goal	-.11	.06	.07	.14
Parents' wishes among reasons most important in deciding to further education	.07	-.07	-.15	-.12

Note. "β" = standardized discriminant function coefficients, directly analogous to regression beta weights. Entries are presented in descending order of the absolute value of the variables' structure coefficients, as regards the function for which variables had the largest $|r_S|$.

The study's third research question focused on what factors predicted students' decisions to change career goals. Seventy-eight of the students reported that they had changed career goals during the last year, a time when as seniors many students become particularly serious in reflecting on their career choices. Table 4 presents a breakdown of the reasons students reported for changing goals. The instrument allowed students the opportunity to select more than one reason. Ten students cited none of the seven alternatives listed on the instrument as a reason for changing their goals. Forty-two students cited one reason; 11 cited two reasons; 11 cited three reasons; and 4 cited four or more reasons.

The same 13 variables used in previous analyses were then employed to predict membership in the group of 124 students who had not changed career objectives, as against the group of 78 students who had. The two groups did not differ to a statistically significant degree ($\lambda = .90$, $\chi^2 = 20.70$, $df = 13$, $p = .08$).

However, a univariate test of one of the 13 interval variables was statistically significant ($F = 5.27$, $df = 1/200$, $p = .02$); this was the univariate test involving felt level of satisfaction since identifying the current career goal. Although this result is noteworthy, it must be remembered that this test was not "protected" by having first found a statistically significant multivariate result, and therefore the result must be interpreted with particular caution.

Table 4
Reasons Cited for Changing Career Goals ($n = 78$)

Reason	n
Job Satisfaction	44
Economic Gain	19
Academic Demands	17
Family Influence	15
Job Prestige	10
Job Stereotype	9
Personal Problems	4

Discussion

The failure to isolate noteworthy differences involving the school sites suggests that the two schools function somewhat similarly as regards recruitment of students and the impacts of curricula. This result is not surprising, since both schools invoke a similar model (cf. Butler et al., 1991).

The failure to find gender differences is encouraging, insofar as the result suggests that equity goals have been realized to at least some degree. Students at these schools make different kinds of career-related choices, but gender does not appear to explain these differences. However, it is important to remember that the students have self-selected into these schools. Some young women, for example, may not aspire to careers in medicine because they have unrealistically low evaluations of personal capacity. Such students would not have thought to apply to one of these schools for the health professions, unless encouraged by a significant other, such as a counselor.

Group differences involving ethnicity, however, had some impact on the interval response variables, as reported in Table 2. The coefficients reported in Table 2 indicate that African-Americans, relative to Hispanics and especially to Asians, were *most* motivated by financial rewards (*F*unction *C*oefficient = -.83; r_s = -.58), were *least* motivated by a desire to "become a more cultured person" (*FC* = +.53; r_s = +.25), felt *least* confidence about being able to meet their career goals (*FC* = +.42; r_s = +.37), and were *most* likely to tie educational objectives more directly to an instrumental effort to obtain career objectives (*FC* = -.14; r_s = -.28).

The African-American students' feelings of less confidence about being able to obtain career objectives may be realistic, especially as regards health careers. There is some evidence (Miller, Thomson, Smith, Thompson & Camacho, 1992) that African-American and Hispanic students do not always receive optimal academic preparation for health careers. Counselors can go a long way toward rectifying deficiencies in the ways that some academic plans have been formulated in the past.

The largest differences involved interest in material rewards; 53% of African-Americans selected being able "to make more money" among the reasons most important in deciding to pursue further education, while 33% of the Hispanics and 24% of the Asians cited this as being an important consideration. This difference does not appear to have resulted from disparate financial situations across the ethnic groups. For example, somewhat similar percentages of African-Americans (33%) and of Asians (24%) indicated that financial obstacles posed the most problems as regards seeking further education. And roughly the same percentages of the African-American students (35%) and the Asian students (39%) were from families in which both parents had obtained a college degree. However, the African-American students may have been from families in which access to financial achievement was a first-generation experience, and consequently, financial achievement may have been seen as both doable and important.

Noteworthy differences were identified also with respect to the career goals selected by the seniors, as reported in Table 3. As reported previously, the centroids (i.e., mean discriminant function scores) on the first discriminant function indicated that this function was most useful for discriminating students selecting career goals in medicine from students selecting other goals. Students selecting medicine as a career goal were *most* confident about their ability to achieve their career objectives (*FC* = +.63; r_s = +.41), were *least* motivated to seek further education "to get a better job" (*FC* = -.60; r_s = -.45), were *most* motivated to seek further education to prepare for graduate school (*FC* = +.49; r_s = +.27), were *least* motivated to seek further education "to get away from home" (*FC* = -.44; r_s = -.44), and were *least* motivated to seek further education so that they could "make more money" (*FC* = -.26; r_s = -.32).

Function II was most useful in distinguishing persons choosing health careers other than medicine from students choosing non-health career goals, as indicated by the group centroids on this function. Students selecting non-medicine health career goals were *most* concerned about seeking further education for the purpose of improving reading and study skills (*FC* = +.51; r_s = +.41), were *least* motivated to seek further education "to become a more cultured person" (*FC* = -.53; r_s = -.35), and were *most* concerned about financial obstacles as regards further education (*FC* = +.56; r_s = +.31).

This profile suggests a continuum with one group of students who are interested in health careers, but have serious reservations about their academic and financial resources; these students decline to abandon the health career interests that presumably first motivated them to enter these schools, but perceive that they have limited options. The students at the other end of this continuum, on the other hand, move toward career objectives in other fields.

Finally, differences involving prediction of changing career goals did not involve a statistically significant multivariate effect. However, the statistically significant, though "unprotected", univariate effect for "what level of satisfaction have you felt since identifying your current career goal" suggests that job satisfaction concerns are important in students' deliberations about career choice. This view is supported by the finding, reported in Table 4, that 44 of 78 students who actually did change career goals during their senior year cited job satisfaction as one of the reasons for the change.

Students came to these schools because of their interest in health careers. A substantial number (78/202 = 39%) changed their career goals while they were enrolled. As indicated by the Table 4 results, some students changed objectives because of perceived economic benefits (*n* = 19), academic demands (*n* = 17), or other influences. But the schools' curricula and professionals apparently did afford students the opportunity to learn about health careers and to make more informed choices about whether such careers will satisfy their needs and interests.

Of course, like all studies, the present study is limited. No one study, taken singly, establishes the basis for generalizable insight (Neale & Liebert, 1986, p. 290). The present study is but one snapshot of dynamics involving the perceptions of minority high school students as regards health-related career choices. Notwithstanding this limitation, the results suggest at least the following conclusions regarding counselors and curriculum developers designing motivational appeals to students or helping minority students to reality-test their expectations.

1. Counselors and curriculum developers may need to pay particular attention to issues involving the perceived financial rewards that some minority students associate with the selection of health careers.
2. Counselors and curriculum developers should note that minority students most interested in careers as physicians differ from minority students interested in allied health or non-health careers, as regards their attitudes and perceptions.
3. Counselors and curriculum developers should attend most closely to job satisfaction issues when facilitating student consideration of changes in career objectives.

Developing multiple snapshots of the career choice dynamics of minority students will enable counselors and curriculum developers better to facilitate informed student choices. The finding that most of these minority students retained an interest in health careers (39% of them in medicine itself) suggests that such programs can be effective in helping to improve the representation of minorities within the health professions.

References

Association of American Medical Colleges. (1992). *Technical assistance manual for Project 3000 by 2000: Guidelines for action.* Washington, DC: Author.

Astin, A .W., Dey, E. L., Korn, W. S., & Riggs, E. R. (1991). *The American freshman: National norms for fall, 1991.* Los Angeles: Higher Education Research Institute, UCLA.

Butler, W. T., Thomson, W. A., Morrissey, C. T., Miller, L. M., & Smith, Q. W. (1991). Baylor's program to attract minority students and others to science and mathematics. *Academic Medicine, 66,* 305-311.

Committee on Allied Health Education and Accreditation. (1991). *Allied health education directory* (19th ed.). Chicago: American Medical Association.

Garcia, R. A., & Ramon, G. (1988). Hispanic enrollment trends in the health professions: Rethinking the problems and solutions. In *Hispanic health status symposium.* San Antonio: Center for Health Policy Development.

Health Resources and Services Administration. (1984). *Minorities and women in the health fields* (DHHS Publication No. (HRSA) HRS-DV 84-5). Washington, DC: U.S. Government Printing Office.

Health Resources and Services Administration. (1990). *Minorities and women in the health fields* (DHHS Publication No. HRSA-P-DV 90-1). Washington, DC: U.S. Government Printing Office.

Huberty, C. J., & Wisenbaker, J. M. (1992). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp. 169-208). Greenwich, CT: JAI Press.

Institute of Medicine. (1989). *Allied health services: Avoiding crisis.* Washington, DC: National Academy Press.

Miller, L. M., LaVois, C., & Thomson, W. A. (1991). Middle school and medical school collaboration: A magnet school experience. *School Science and Mathematics, 91*(2), 47-50.

Miller, L. A., Thomson, W. A., Smith, Q., Thompson, B., & Camacho, Z. (1992). Perceived barriers to careers involving math and science: The perspective of medical school admissions officials. *Teaching and Learning in Medicine, 4*, 9-14.

Mingle, J. R. (1987). *Focus on minorities: Trends in higher education participation and success.* Washington, DC: Education Commission on Status and State of Higher Education.

Neale, J. M., & Liebert, R. M. (1986). *Science and behavior: An introduction to methods of research* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Shugars, D. A., O'Neil, E. H., & Bader, J. D. (Eds.). (1991). *Healthy America: Practitioners for 2005, an agenda for action for U.S. health professional schools.* Durham, NC: The Pew Health Commission.

Super, D. E. (1951). Vocational adjustment: Implementing a self-concept. *Occupations, 30*, 88-92.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement, 45*, 203-209.

Thomson, W. A., Holcomb, J. D., & Miller, L. M. (1987). Enhancing career opportunities for students in the health professions through high school apprentice program. *Journal of Studies in Technical Careers, 9*, 113-124.

Thomson, W. A., Smith, Q. W., Miller, L. M., & Shargey, B. O. (1991). Survey of 1975-1987 graduates of Houston's High School for the Health Professions. *Academic Medicine, 66*, 364-366.

U.S. Department of Health and Human Services. (1985). *Report of the secretary's Task Force on Black and Minority Health.* Washington, DC: U.S. Government Printing Service.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 29-35

Leadership for Productive Schools

William L. Johnson, Karolyn J. Snyder, and Annabel M. Johnson

The expanding research literature based on productive industries, social agencies, and effective schools identifies exemplary production patterns found in dynamic work cultures. In a study conducted in two large school districts in the United States, educational administrators expressed a strong desire for training in the major areas of educational leadership that were assessed: The principalship, problem solving, planning for school growth, personal awareness, staff development, long-range planning, and the school as a system. These findings have numerous implications pertaining to school-based educational leadership.

The management writings that have made the best-seller lists in recent years (e.g., Deal & Kennedy, 1982; Geneen with Moscow, 1984; Kanter, 1983; Naisbett, 1982; Peters, 1988; Peters & Waterman, 1982), the generic base of management and organizational theory and research, and the studies of effective schools have all pointed to the critical role of the school principal and the principal's potential ability to alter work and achievement patterns. Accordingly, the vital function of school-based leadership is being examined by persons in departments of education, professional organizations, and school districts and by principals themselves.

Even though the research literature corresponds with our intuitions about good schools and effective leadership (Purkey & Smith, 1985), error potential resides in merely exhorting principals to focus on school-based leadership and go forth and lead. In fact, most principals today are simply not prepared to meet the school's need for instructional leadership. Before 1950, principals focused primarily on being instructional leaders. However, in post-World War II America, as schools grew larger and more complex, administrators' emphasis swung toward personnel, budget, and public relations (Goodland, 1979).

This view of the educational administration as the "corporate" chief executive officer has been modified by at least three recent trends (Bryant, 1988; Deming, 1986): (a) restructuring schools, (b) recognizing leadership styles and differences (Statham, 1987), and (c) understanding the ecology of school improvement (Eisner, 1988). These and other trends are reflected in the research base that undergirds productive organizations.

Numerous studies in recent years confirm that strong instructional leaders are critical factors in effective schools. Strong principal instructional leadership has been shown to be correlated with school effectiveness (Bossert, Dwyer, Roward, & Lee, 1982; Hallinger & Murphy, 1986). A Rand study of 1977 called the principal the "gatekeeper" of change and reported that principals were powerful enough to prevent and foster any kind of change within their schools. Additionally, DeBevoise (1984), Hallinger and Murphy (1986), and Larsen (1989) concluded from their studies that instructional leadership was indeed a key to an effective school.

But how is a principal's behavior steered in the direction of instructional leadership? Outlining conceptual or methodological guidelines for accomplishing instructional leadership is especially difficult since little consensus exists among school administrators who discuss these tasks. Moreover, how do these tasks differ from what principals have always done? Furthermore, researchers have rarely defined instructional leadership in terms of specific policies, practices, and behaviors initiated by the principal (Hallinger & Murphy, 1987; Larsen, 1987).

The purpose of our paper is threefold. *First*, we will synthesize the results of effective schooling characteristics and leadership tasks that have been identified by the research community and use this research base of over 400 research studies as a focus to present a production model for instructional leadership. *Second*, we will report the instructional leadership training needs for administrators that we studied in two large school districts in the United States. Our pool of respondents ($n = 279$) for the

---

William L. Johnson is a Professor, Chair of the Psychology and Education Department, and Associate Dean of Academic Affairs at Ambassador College in Big Sandy, Texas. Karolyn J. Snyder is a Professor of Education at the University of South Florida, Tampa, Florida, where she directs the School Management Institute. Annabel M. Johnson is a Professor of Home Economics at Ambassador College in Big Sandy, Texas. Her specialty is in the area of management. Please address correspondence regarding the paper to William L. Johnson, Associate Dean of Academic Affairs, Ambassador College, Big Sandy, TX 75755.

study were elementary and secondary school principals plus central office supervisory personnel.

Research on Effective School Characteristics

Overall Organizational Culture

Culture is so powerful a force in organizations that it either stimulates or represses competent performance (Kanter, 1983). As a powerful, intangible force, it functions to limit or to enhance the capacity of an organization to respond to issues (Carlopio, 1988). Culture includes the physical and social structure of an organization as well as the values and assumptions of individuals within the organization. Culture is more than climate (how people feel about an organization): It is the social and psychological force that stimulates the direction and quality of work in an organization (Snyder, 1988). Schein (1985) suggests that "culture and leadership are two sides of the same coin" and contends that "the only thing of real importance that leaders do is to create and manage culture" (p. 2).

"Work culture" is conceptualized in the next several paragraphs to include four interdependent dimensions: structures and processes for school planning, professional systems and tools, program development processes, and school assessment systems. Together these dimensions provide the direction and energy system for those in a school (or other organization) to alter the organization's programs and structures and to enhance its efforts upon learning patterns.

Dimension 1: School-Wide Planning

Administrators and employees together transform common concerns into specific achievement-oriented development goals. Planning tasks include setting organizational goals that relate to primary outcomes and visions for the organization (Conley, Schmidel, & Shedd, 1988; Davidson & Montgomery, 1985). Next, tasks are dispersed to a variety of permanent and ad hoc work groups that function collaboratively, forming and reforming as needs are addressed (Cook, 1982; Deal & Kennedy, 1982). Individuals are held accountable for their contributions within small work units (Drucker, 1982; Levin, 1986). Peters and Austin (1985) found that the intensity of management's commitment to organizational goals was the chief difference between great and not-so-great organizations.

Dimension 2: Professional Development

Professional development plans that are linked to organizational goals have the power to enhance individual and group performance, and that of the school as well (Carneval, 1989; Glenn, 1981). Managers and workers regularly coach each other as they develop new skills and solve problems (Clark, 1985; Garmston, 1987). Work groups become learning centers for teachers as they share, plan, act, and critique programs or tasks together (Larson & LaFasto, 1989; Little, 1982). Collaborative quality control systems are replacing outdated monitoring systems and provide for regular group reflection, data analysis and problem solving as the organization works on its plans (Peters & Waterman, 1982). Quality control in the best institutions today is viewed as developmental and provides opportunities for work adjustment in fast paced, changing environments (Wise & Darling-Hammond, 1984).

Dimension 3: Program Development

Principals and supervisors convey instructional standards to workers in productive schools (Coulsen, 1977). They also coordinate program development, implementation, and testing activity to stimulate change in order to address learning challenges (Venezky & Winfield, 1979). The purpose of managing program development in the best schools is to solve specific problems, doing whatever it takes to solve learning challenges (Austin, 1979). It is also well documented that high levels of parent and community involvement facilitate student success patterns. This kind of involvement is subsumed under the perceptions of the principal's governance of the school's instructional program (Heck, Larsen, & Marcoulides, 1990).

Dimension 4: School Assessment

Accountability systems drive assessment activity in productive organizations (Brookover & Lezotte, 1979). The only assessment system that appears to have the power to alter individual and organizational performance is a goal-based system (Odiorne, 1979). When organization members assess how well they have achieved their goals and analyze data from each level of work, they learn that goals provide a word focus that leads to organizational success (McGregor, 1960). Assessment data in productive organizations provide both a feedback and a feed-forward loop that influence both short and long-range planning (Michael, Luthans, Warner, & Hayden, 1981).

The expansion of the knowledge base about organizational and human productivity over the past decade indicates that until administrators and teachers together assume responsibility for schooling, achievement patterns are likely to remain unchanged. These conclusions support a model of multidimensional managerial leadership behavior within the school context. Interestingly, Pitner and Hocevar (1987) applied confirmatory factor analysis to Yukl's (1981) multidimensional leadership model and identified 14 domains of principal leadership behavior. We believe that the management challenge is to empower groups to address educational productivity. The administrative challenge is one of instructional leadership and restructuring (integrating teachers into the decision-making process of the schools). This synthesis emerges from the developing work-culture literature (Carlopio, 1988; Kanter, 1983; Kilmann, Corwin, & Associates, 1988; Levinson, 1968; Selman & DiBianca, 1983) and the instructional leadership findings mentioned previously.

Indeed, the literature on all kinds of productive organizations continues to affirm clearly and strongly that employee involvement in shaping the organization's direction is essential to the very survival of an organization. Resources, information, and opportunity are the vital materials that fuel organizational productivity (Johnson & Snyder, 1989-1990). A typical production model might divide the school year into three parts: planning (September and October), development (November through April), and evaluation (May and June). Planning activities would include school-wide goals setting and work group and individual staff performance planning. Developmental activities might include staff development, clinical supervision, work group development, and quality control activities. Program development might include instructional program and resources development. School productivity assessment would include assessing achievement for students, teachers, work groups, and the school itself. The assessment findings would then serve to direct the feedback and feed-forward planning and development activities for the next academic year.

Methodology

Once the literature was assessed, considering this empirical base of over 400 research studies of productive organizations (75%) and effective schools (25%) (Snyder & Anderson, 1986; Snyder, 1988), we sought to measure the existing training needs for principals and district personnel vis-a-vis the model. Consequently, a 76-question needs assessment instrument was developed. In the 1980s, the authors surveyed about 450 school administrators in eight school districts in the United States and used the data to validate the instructional leadership needs assessment instrument (Johnson & Snyder, 1986; Johnson, Snyder, & Johnson, 1992-1993). The articles discuss the development of the instrument and give the psychometric properties of such. Table 1 indicates that 50 items remained in the instrument following the elimination of 26 items suggested as inappropriate by the principal factor analysis.

Table 1
Psychometric Properties of the Instrument

Item Number	Category	Factor Loading	Mean	Standard Deviation
1	Goal Setting	0.536	3.65	1.44
2	Goal Setting	0.422	3.80	1.43
3	Goal Setting	0.534	3.75	1.51
4	Goal Setting	0.464	3.91	1.48
5	Goal Setting	0.489	3.46	1.45
6	Goal Setting	0.464	3.73	1.51
7	School as an Ecosystem	0.405	3.86	1.40
8	School as an Ecosystem	0.713	3.44	1.38
9	School as an Ecosystem	0.711	3.26	1.42
10	Problem Solving	0.643	3.94	1.36
11	Problem Solving	0.676	4.09	1.37
12	Problem Solving	0.461	3.90	1.31
13	Problem Solving	0.455	4.14	1.34
14	Principalship	0.413	3.86	1.37
15	Problem Solving	0.401	4.10	1.40
16	School as an Ecosystem	0.362	3.46	1.43
17	Principalship	0.446	3.95	1.47
18	Principalship	0.401	4.01	1.36
19	Planning	0.441	3.89	1.38
20	Planning	0.424	3.93	1.33
21	Staff Development	0.414	3.83	1.40
22	Principalship	0.564	3.90	1.35
23	Planning	0.564	3.70	1.43
24	Principalship	0.521	3.92	1.34
25	School as an Ecosystem	0.305	3.65	1.50
26	Personal Awareness	0.738	3.90	1.38
27	Personal Awareness	0.731	3.92	1.36
28	Planning	0.489	3.91	1.43
29	Planning	0.548	3.87	1.43
30	Planning	0.408	3.90	1.45
31	Planning	0.534	3.87	1.43
32	Personal Awareness	0.403	3.94	1.34
33	Planning	0.471	3.74	1.43
34	Planning	0.430	3.86	1.33
35	Planning	0.450	3.84	1.47
36	Personal Awareness	0.645	4.07	1.37
37	Planning	0.500	3.72	1.51
38	Planning	0.521	3.75	1.48
39	Planning	0.655	4.03	1.38
40	Planning	0.540	4.03	1.37
41	Staff Development	0.718	4.03	1.37
42	Staff Development	0.463	3.98	2.32
43	Staff Development	0.717	4.13	1.34
44	Planning	0.407	3.86	1.42
45	Personal Awareness	0.692	3.82	1.38
46	Planning	0.559	3.70	1.49
47	Personal Awareness	0.522	3.91	1.37
48	Planning	0.559	3.95	1.48
49	Planning	0.570	3.69	1.44
50	Staff Development	0.466	4.16	1.50

BEST COPY AVAILABLE

The authors designed a one column quasi-Likert response format for the instrument. The educational leaders were asked to select from a range of six levels of need: "1" no training (skill unrelated); "2" no training (competency high); "3" training (awareness level); "4" training (initial practice); "5" training (skill refinement); and "6" assistance with school implementation. Based on our literature review, the authors examined seven areas that were rated highly as foci for leadership development: the principalship, the school as a system, problem solving, staff development, long range planning, goal setting, and personal awareness. The questions from all seven categories were then randomly assigned to the survey instrument. See Table 2 for a representative question from each category of the instrument. The Cronbach alphas for the categories had a range from .76 to .86.

Table 2
Representative Questions from the Survey

Category	Item
Problem Solving	Leading a staff toward creative solutions to problems
Staff Development	Evaluating teacher performance
The Principalship	Developing strategies for supervisory conference feedback
Personal Awareness	Assessing my own professional growth needs
Long Range Planning	Developing strategies for accomplishing school goals
Goal Setting	Developing a school-wide goal setting process
School as a System	Identifying the subsystems of school (district) organization and their effects on school performance

The literature review and brief description of the instrument provide the context for the present study. Both show that the present study is based on a strong research foundation. The respondents for this study were 279 district administrators and central office personnel in two large school districts in the United States. The respondents were part of the national sample that was surveyed to validate the administrative needs assessment instrument. For the first district, the respondents represented 70% of the district's administration personnel. For the second district, all 76 elementary principals responded along with 8 elementary assistant principals. Overall there were 151 elementary principals and assistant principals, 64 secondary principals and assistant principals, plus 64 central office supervisory personnel.

Regarding demographic information, the respondents were asked about their job title, division for which they were responsible, the size of their school district, and the setting of their district. Two hundred and thirty-four of the respondents' schools were in urban areas. The other 45 respondents' schools were in suburban areas. Fifty-four respondents were in districts with less than 500 scholastics, 140 had from 500 to 2,000 scholastics, and 85 were from districts with more than 2,000 scholastics.

Findings and Discussion

The elementary and secondary school principals and central office personnel who responded to the instrument reported that they desired training in all seven categories, with the desires ranging between training at the awareness level (category 3) and at the initial practice level (category 4). We interpreted the range of scores to indicate that new knowledge and skills in all categories were perceived as important to the administrators' role success. A category response range of 1 (low) to 6 (high) was possible. Table 3 outlines the findings of the study.

Problem solving was rated most highly, reflecting the dramatic changes in job expectations and the dynamic work culture of the schools. This category addressed techniques and processes that can be used in solving real school problems in a collaborative mode. Staff development was rated second. This category focused on ways to develop and operationalize a school program for staff growth that emphasizes new knowledge and skills that are necessary for successful attainment of school development goals (school, work, and individual). The principalship was rated next. The principalship category consisted of questions pertaining to school leadership and organization, staff motivation, and directing school activities. Last in rank was the school as a system which was the category describing the school's ecology and the many organizational factors which work interdependently to influence achievement results. Overall, this category addressed questions relating to environmental factors, such as federal, state, community, parental, and district pressures, and factors that are internal to the staff, such as students, programs, achievement levels, and staff competency. While all seven categories are distinct from each other, each seemed to represent and be of concern to principals in providing effective instructional leadership. Our observations suggest, moreover, that principals do not exercise these content areas to an equal extent across all seven areas. Assessment of the extent of the exercise of these areas would comprise another important study.

Table 3
Ranking (High to Low) of the Need Indices for the Instructional Leadership Surveys
(All Educational Leaders n = 279)

Rank	Area Index	Need	Standard Deviation
1	**Problem Solving** (questions relating to cooperative decision-making)	3.98	1.44
2	**Staff Development** (questions relating to developing a school program for staff growth)	3.93	1.47
3	**The Principalship** (questions relating to instructional leadership expectations)	3.88	1.46
4	**Personal Awareness** (questions relating to the leader's self-concept, personality, leadership style, and their influence on instructional leadership behaviors)	3.87	1.45
5	**Long-Range Planning** (questions relating to cooperative action planning, monitoring, and evaluation)	3.83	1.50
6	**Goal Setting** (questions relating to organizational analysis and goal setting for school leadership)	3.70	1.55
7	**School as a System** (questions relating to school goals, organization, performance, program, technology and management, how together these guide the school improvement process)	3.53	1.47

Those studied perceive that the categories surveyed and identified in the research literature are important in their job and also are a desirable focus for their own professional development. Furthermore, the skills necessary for successful collaboration, organizational assessment and analysis and a knowledge of how personal characteristics influence leadership appear to be important to the elementary and secondary administrators for the successful implementation of instructional leadership tasks.

Pertaining to the limitations of this study, the authors have reported the need indices to two decimal places, although the data were collected as integers. Indeed, this degree of computation may represent "over-scientificism." The researchers conclude that those surveyed perceive all the categories important in their jobs and also a desirable focus for their own professional development.

School districts, professional organizations, and university professional programs need to devote priority attention to all these needs. School district leaders need to augment workshops for their teachers and administrators. Furthermore, peer supervision and coaching, and peer efforts to develop and implement new skills in the classroom benefit the school system. A feedback mechanism of some sort is also essential to the eventful successful development of expertise in instructional leadership.

Summary and Conclusion

Educational administration is changing from an emphasis on just administering policy and managing compliance to a focus on leading instructional improvement efforts. Principals are expressing a desire for the skills necessary to become successful instructional leaders. Our study of principals reinforces our observations: Because of a redefinition of the principalship, principals themselves are faced with a need for new job knowledge and skills. Further, principals want training in the elements of annual school-wide, team-level, and individual teaching planning, coaching, and evaluation. In addition, they want skills for designing successful staff development programs, providing on-the-job teacher coaching, monitoring performance and program development, implementation, and evaluation. Moreover, in addition to the tasks of instructional leadership, principals also want to know how to involve others successfully in cooperative planning and action. Furthermore, the participants report that there is a major concern for motivating teachers to work in more productive ways. We note that supervision of the school's instructional organization is perceived as a major component in the principals' instructional leadership activities. This would include activities related to monitoring teacher performance. This concern would include, but not be limited to, such activities as establishing school goals or identifying inservice needs.

The message for role development is clear: If principals are expected to perform new tasks and accomplish different kinds of performance results from that for which they were educated, their development in a new set of knowledge and skills must become a district priority. Moreover, our findings in this study lead us to challenge school administrators to foster skill development by combining initial training in instructional leadership tasks with continuous on-the-job peer and supervisory coaching.

References

Austin, G. R. (1979). Exemplary schools and the search for effectiveness. *Educational Leadership, 37*(2), 10-14.

Bossert, S., Dwyer, D., Rowan, B., & Lee, G. (1982). The instructional management role of the principal. *Educational Administration Quarterly, 18*(3), 34-36.

Brookover, W. B., & Lezotte, L. W. (1979). *Changes in school characteristics coincident with changes in student achievement.* East Lansing, MI: Institute for Research on Teaching, College of Education, Michigan State University. (ERIC Document Reproduction Service No. ED 181 005)

Bryant, M. T. (1988). An inquiry-based orientation for preparation programs in educational administration. *National FORUM of Educational Administration and Supervision Journal, 6*(1), 70-80.

Carlopio, J. (1988). Make a difference: Break through your organizational limits. *Organizational Development Journal, 6*(3), 46-50.

Carneval, A. P. (1989). The learning enterprise. *Training and Development Journal, 43*(2), 26-33.

Clark, C. (1985). *A comprehensive program of evaluation and professional development: A working model.* (ERIC Document Reproduction Service No. ED 270 866)

Conley, S. C., Schmidel, T., & Shedd, J. B. (1988, Winter). Teachers' participation in the management of school systems. *Teachers College Record, 90*(2), 260-280.

Cook, M. H. (1982). Quality circles--they really work, but..." *Training and Development Journal, 36*(1), 4-6.

Coulsen, J. E. (1977). *Overview of the National Evaluation of the Emergency School Aid Act.* Santa Monica, CA: System Development Corporation. (ERIC Document Reproduction Service No. ED 154 951)

Davidson, J. L., & Montgomery, M. (1985). *Instructional leadership system research report.* Paper presented at the Annual Conference of the American Association of School Administrators, Dallas, TX.

Deal, T. E., & Kennedy, A. A. (1982). *Corporate cultures: The rites and rituals of corporate life.* Reading, MA: Addison-Wesley.

DeBevoise, W. (1984). Synthesis of research on the principal as instructional leader. *Educational Leadership, 41*(5), 14-20.

Deming, W. E. (1986). *Out of the crisis* (2nd ed.). Boston: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Drucker, P. F. (1982). *The changing world of the executive.* New York: Truman Tally Books.

Eisner, E. W. (1988). The ecology of school improvement. *Educational Leadership, 45*(5), 24-29.

Garmston, R. J. (1987). How administrators support peer coaching. *Educational Leadership, 44*(5), 18-26.

Geneen, H., with Moscow, A. (1984). *Managing.* Garden City, NJ: Doubleday.

Glenn, B. C. (1981). *What works? An examination of effective schools for poor black children.* Cambridge, MA: Harvard University, Center for Law and Education.

Goodland, J. I. (1979). *What are schools for?* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Hallinger, P., & Murphy, J. (1986). The social context of effective schools. *American Journal of Education, 94*(3), 328-355.

Hallinger, P., & Murphy, J. (1987). Instructional leadership in the school context. In W. Greenfield (Ed.), *Instructional leadership: Concepts, issues, and controversies* (pp. 179-203). Boston: Allyn & Bacon.

Heck, R. H., Larsen, T. J., & Marcoulides, G. A. (1990). Instructional leadership and school achievement: Validation of a model. *Educational Administration Quarterly, 26*(2), 94-125.

Johnson, W. L., & Snyder, K. J. (1986). Instructional leadership effectiveness: A research analysis and strategy. *Educational and Psychological Research, 6*(1), 27-47.

Johnson, W. L., & Snyder, K. J. (1989-1990). Planning for faculty development in America's colleges. *National FORUM of Applied Educational Research Journal, 2*(1), 34-38.

Johnson, W. L., Snyder, K. J., & Johnson, A. B. (1992-1993). Developing instruments for educational administration. *National FORUM of Applied Educational Research Journal, 6*(1), 3-11.

Kanter, R. M. (1983). *The change masters: Innovation for productivity in the American corporation.* New York: Simon & Schuster.

Kilmann, R. H., Corwin, T., & Associates. (1988). *Corporate transformation: Revitalizing organizations for a competitive world.* San Francisco: Jossey-Bass.

Larsen, T. (1987, April). *Synopsis: Identification of instructional leadership behaviors and the impact of their implementation on academic achievement.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Larsen, T. (1989). Effective schools? Effective principals! *Thrust for Educational Leadership, 19*(1), 35-36.

Larson, C. E., & LaFasto, F. M. J. (1989). *Team work: What must go right/what can go wrong.* Newbury Park, CA: Sage Publications.

Levin, H. Z. (1986). The squeeze on middle management. *Personnel, 63*(1), 62-69.

Levinson, H. (1968). *The exceptional executive: A psychological conception.* Cambridge, MA: Harvard University Press.

Little, J. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal, 19*(3), 325-340.

McGregor, D. (1960). *The human side of the enterprise.* New York: McGraw-Hill.

Michael, S. R., Luthans, F. S., Warner, G. W., & Hayden, S. (1981). *Techniques of organizational change.* New York: McGraw-Hill.

Naisbett, J. (1982). *Megatrends: Ten new directions transforming our lives.* New York: Warner Books.

Odiorne, G. S. (1979). *MBO II: A system of managerial leadership in the 80s.* Belmont, CA: Fearon Pittman.

Peters, T. J. (1988). *Thriving on chaos: A handbook for a management revolution.* New York: Alfred A. Knopf.

Peters, T. J., & Austin, N. A. (1985). *A passion for excellence: The leadership difference.* New York: Random House.

Peters, T. J., & Waterman, R. H. (1982). *In search of excellence: Lessons from America's best run companies.* New York: Harper & Row.

Pitner, N., & Hocevar, D. (1987). An empirical comparison of two-factor versus multifactor theories of principal leadership: Implications for the evaluation of school principals. *Journal of Personnel Evaluation in Education, 1*(1), 93-109.

Purkey, S. C., & Smith, M. S. (1985, June). *Effective schools: A review.* Madison, WI: Wisconsin Center for Educational Research.

Schein, E. H. (1985). *Organizational culture and leadership.* San Francisco: Jossey-Bass.

Selman, J., & DiBianca, V. E. (1983). Contextual management: Applying the art of dealing creatively with change. *Management Review, 72*(9), 13-19.

Snyder, K. J. (1988). *Managing productive schools.* San Diego: Harcourt Brace & Jovanovich.

Snyder, K. J., & Anderson, R. H. (1986). *Managing productive schools: Toward an ecology.* San Diego: Harcourt Brace & Jovanovich.

Statham, A. (1987). The gender model revisited: Differences in the management styles of men and women. *Sex Roles: A Journal of Research, 16*(7/8), 409-426.

Venezky, R. L., & Winfield, L. F. (1979). *Schools that succeed beyond expectation in reading.* (Studies on Educational Technology, Report No. 1). Newark, DL: University of Delaware.

Wise, A. E., & Darling-Hammond, L. (1984). Teacher evaluation and teacher professionalism. *Educational Leadership, 42*(4), 28-33.

Yukl, G. (1981). *Leadership in organization.* Englewood Cliffs, NJ: Prentice-Hall.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 37-47

The Relationship of the Murphy-Meisgeier Type Indicator for Children to Sex, Race, and Fluid-Crystallized Intelligence on the KAIT at Ages 11 to 15

Alan S. Kaufman and James E. McLean

Four typologies assessed by the Murphy-Meisgeier Type Indicator for Children (Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, Judging-Perceiving) were related to sex, race/ethnic group, intelligence level, and Fluid/Crystallized IQ discrepancy. IQ scores were obtained using the Kaufman Adolescent and Adult Intelligence Test (KAIT). Data from 263 individuals aged 11 to 15 years were subjected to MANOVAs and MANCOVAs, covarying parents' education. No interactions were significant, and sex was the only significant main effect. Univariate ANOVAs and ANCOVAs indicated that the Thinking-Feeling index produced the significant main effect, a finding consistent with previous research (females favored Feeling more so than males). Educational implications of the findings are provided.

The Murphy-Meisgeier Type Indicator for Children (Meisgeier & Murphy, 1987) is a downward extension of the Myers-Briggs Type Indicator (Briggs & Myers, 1983; Myers & McCaulley, 1985), a widely used personality test; both instruments are derived from Jung's theory of psychological types. The Myers-Briggs has been typified as "an excellent example of a construct-oriented test" (Wiggins, 1989, p. 538) and "probably the most widely used instrument for non-psychiatric populations in the areas of clinical, counseling, and personality testing" (DeVito, 1985, p. 1030). Meisgeier and Murphy (1987) note the wide and diverse use of the Myers-Briggs for counseling, career planning, staff and professional development, education, and personal growth, and state: "A similar approach describing individual differences is at least as important [for children], and possibly much more so. With children, the issues are related not only to understanding the type of oneself and others, but also to the development of type in healthy and functional ways" (p.1). With the latter position serving as Meisgeier and Murphy's (1987) rationale, and in view of their perception that the "means to identify psychological type in children have been practically nonexistent" (p. 1), they constructed the Murphy-Meisgeier Type Inventory for Children, Form D. The instrument is intended primarily for children in grades 2 through 8, or approximately ages 7 to 15 years.

The Murphy-Meisgeier provides scores on the same four Jung-inspired indices as the Myers-Briggs: Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, and Judging-Perceiving. But whereas research has been plentiful on the Myers-Briggs (e.g., Carlson, 1985, 1989; Dilley, 1987; Lynch, 1985; Myers & McCaulley, 1985), empirical studies have been notably lacking on the Murphy-Meisgeier. The children's inventory has advocates (Allen, 1989), but systematic investigation of its validity seems to have been limited to concurrent validity and canonical correlation studies reported in Murphy's (1986) doctoral dissertation, in the manual (Meisgeier & Murphy, 1987), and in an article (Fourqurean, Meisgeier, & Swank, 1990). Criteria have been measures of personality and learning styles; the implications of the results of these studies are of limited value for the counselor's and psychologist's understanding of the meaning of the test scores.

The aim of this investigation was to study the Murphy-Meisgeier for a sample of children ages 11 to 15 years who were tested during the nationwide standardization of the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993). Just as the Murphy-Meisgeier is a downward extension of the Myers-Briggs, this study is a downward extension of investigations that related the Myers-Briggs to the KAIT for individuals aged 14 to 94 years (Kaufman, Kaufman, & McLean, 1993; Kaufman, McLean, & Underwood, 1992, November). Those studies related Jungian type to age, sex, race/ethnic group, and fluid and crystallized intelligence on the KAIT. Results indicated that: (a) Age

---

Alan S. Kaufman is a Research Professor of Behavioral Studies in the College of Education at The University of Alabama. James E. McLean is a University Research Professor and the Assistant Dean for Research and Service in the College of Education at The University of Alabama. The authors wish to express their appreciation to Vicki Benson who shared her insights into the implications for classroom teachers and administrators of the results of the study and to Sherry Pasquale who shared an unpublished paper regarding gender equity in education. Please address correspondence regarding the paper to James E. McLean, Office of Research and Service, The University of Alabama, Tuscaloosa, AL 35487-0231 (Internet: JMCLEAN@UA1VM.UA.EDU).

was significantly related to Judging-Perceiving, with younger people tending to be more Perceiving and middle-aged and elderly individuals tending to be more Judging; (b) Sex was significantly related to Thinking-Feeling, with females tending to score at the Feeling end of the continuum and males tending to score at the Thinking end; (c) Race/ethnic group also was significantly related to Thinking-Feeling, with African-Americans favoring a Thinking decision-making style, and both Anglo-Americans and Hispanics split about evenly between the two styles; (d) IQ level was significantly related to Sensing-Intuition, with low functioning and average ability individuals favoring Sensing as a means of receiving information, and high functioning people using the two functions for receiving information about equally; and (e) the discrepancy between fluid and crystallized intelligence did not relate significantly to any typology (Kaufman et al., 1993; Kaufman et al., 1992, November).

In the present study, Sex, Race/ethnic group, IQ level, and Fluid-Crystallized discrepancy were investigated, but chronological age was excluded from the multivariate analyses because (a) the sample was quite limited in its age range, and (b) its inclusion would have resulted in cell sizes of zero. Nonetheless, the variable of age was investigated informally by comparing the results for 11-15 year-olds on the Murphy-Meisgeier with the Myers-Briggs results for ages 14-19, 20-29, 30-49, and 50-94 (Kaufman et al., 1992, November).

The significant Sex main effect for the Thinking-Feeling dimension in the Kaufman et al. (1992, November) study is consistent with many other Myers-Briggs findings (Myers & McCaulley, 1985), and a similar result holds for the Murphy-Meisgeier (Meisgeier & Murphy, 1987, Table 12). Consequently, significant female-male differences on Thinking-Feeling were hypothesized for the present sample of 11-15 year-olds.

On the Myers-Briggs, Anglo-Americans and Hispanics displayed similar typologies, but both ethnic groups differed from African-Americans on the Thinking-Feeling dimension (Kaufman et al., 1993). The Thinking-greater-than-Feeling pattern on the Myers-Briggs for African-Americans was obtained for a group of high school students (Melear & Pitchford, 1991), but not for a group of college students (Levy, Murphy, & Carlson, 1972). Nonetheless, the latter investigation did reveal dramatic ethnic differences when African-American college students were compared to Anglo-American college students, with the African-American students evidencing a predominance of Sensing and Judging types, relative to the Anglo-American students. Consequently, significant Race/ethnic differences were anticipated in the present study, especially for African-Americans versus Anglo-Americans, on one or all of the following indices: Thinking-Feeling, Sensing-Intuition, Judging-Perceiving.

The relationship of high intelligence to relatively high scores on Myers-Briggs Intuition has been demonstrated with numerous group tests such as the Scholastic Aptitude Test (SAT) for numerous adolescent and young adult samples (Myers & McCaulley, 1985). The Kaufman et al. (1992, November) study extended that finding to the individually administered, clinically-oriented KAIT, and to a broad age range from early adolescence to old age. The significant relationship of KAIT IQ level to Sensing-Intuition was observed for each age group in the Kaufman et al. investigation (1992, November), and a significant relationship was therefore hypothesized for the 11-15 year-olds in the present sample.

The KAIT provides measurement of fluid intelligence, the ability to solve novel problems that are not school taught, and of crystallized intelligence, the ability to answer questions that depend on schooling and acculturation for success. The discrepancy between these two theoretical constructs, derived from the Horn-Cattell theory of intelligence (Horn, 1989; Horn & Cattell, 1966, 1967), was anticipated by Kaufman et al. (1992, November) to relate meaningfully to one or more of the Jungian constructs measured by the Myers-Briggs. The fact that the two sets of constructs did not relate significantly leads to the hypothesis that Fluid-Crystallized discrepancy will not emerge as a significant main effect in the present analysis of the Murphy-Meisgeier.

Method

Subjects

Sample participants were 263 preadolescents and adolescents aged 11 to 15 years (mean = 12.0, SD = 1.0). The group included 122 females (46.4%) and 141 males (53.6%), and was composed of 192 Anglo-Americans (73.0%), 38 African-Americans (14.4%), and 33 Hispanics (12.6%). The small number of African-Americans and Hispanics was a result of the sample mirroring the national population. Mean age was 12.0 (SD = 0.9) for females, 12.0 (SD = 1.0) for males, 12.0 (SD = 0.9) for Anglo-Americans, 12.1 (SD = 1.0) for African-Americans, and 12.1 (SD = 1.0) for Hispanics. Parents' educational attainment was used to estimate socioeconomic status. Mean education for the total sample was 13.6 years of schooling (SD = 2.8). Mean education level was 13.6 (SD = 3.2) for females, 13.6 (SD = 2.5) for males, 14.2 (SD = 2.3) for Anglo-Americans, 13.0 (SD = 2.6) for African-Americans, and 10.9 (SD = 3.8) for

Hispanics. Data also were available on 20 "others" (e g., Native Americans, Asian-Americans), but they were not included in this study because (a) its sample size was too small for the MANOVAs and MANCOVAs that were conducted, and (b) it did not constitute a homogeneous or meaningful Race/ethnic group. Data from two individuals ages 16-17 also were excluded because they were older than the optimal age range proposed by the authors of the Murphy-Meisgeier.

Subjects were tested throughout the United States during the nationwide standardization of the KAIT (Kaufman & Kaufman, 1993).

Instruments

Murphy-Meisgeier Type Indicator for Children, Form D. The Murphy-Meisgeier (Meisgeier & Murphy, 1987) is a self-report inventory that provides scores on the same four separate Jungian indices used by the Myers-Briggs (abbreviations for each preference are shown in parentheses):

1. Extraversion (E)-Introversion (I), designed to reflect whether a person is an extravert or introvert in Jung's sense of these terms: Extraverts relate more easily to the outer world of people and things whereas introverts relate more easily to the inner world of concepts and ideas.

2. Sensing (S)-Intuition (N), designed to reflect a person's preference between two opposite ways of perceiving: sensing (reports observable facts or happenings through one or more of the five senses) versus intuition (reports possibilities and relationships).

3. Thinking (T)-Feeling (F), designed to reflect a person's preference between two opposite ways of judging: thinking (bases judgments on impersonal analysis and logic) versus feeling (bases judgments on personal values).

4. Judgment (J)-Perception (P), designed to reflect a person's preference for dealing with the outer or extraverted world either by judgment or by perception; the one who prefers judgment deals with the outer world by thinking or feeling, whereas the one who prefers perception deals with the outer world by sensing or intuition.

Individuals respond to 70 items that deal with inconsequential everyday events. As for the Myers-Briggs, sets of items were developed for each of the four preference scales; for each item, examinees must make a forced choice between the poles of a particular index. Preference scores for each index are obtained by summing item scores. Each item is weighted based on weights derived from a discriminant analysis procedure. Scores can range from 35 to 70 on E-I; from 44 to 88 on S-N and J-P; and from 42 to 84 on T-F. Low scores indicate preferences for the first-named pole of each scale (E, S, T, or J) and high scores indicate preferences for the second-named pole (I, N, F, or P). Unlike the Myers-Briggs, the Murphy-Meisgeier does not "force" one pole of each index to be the preference. For each scale, there is a "U-Band" that indicates an undetermined preference. These bands of "no preference" are as follows for each index: E-I (47.7-52.3), S-N (64.4-69.6), T-F (61.6-66.4), J-P (63.9-68.1). Preference scores less than the lower bound of the U-Band are coded with the first-named pole of the index; preference scores greater than the upper bound are coded with the second-named pole. Taken together, the four preferences denote the person's "type," and are abbreviated as ENFP, ISTP, INFJ, and so forth, similar to the Myers-Briggs. With the Murphy-Meisgeier, however, types such as INUP or UUTJ are possible, where "U" always denotes an "Undetermined" preference. Both the letter designations and the preference scores indicate the direction of the preference, but neither one denotes the magnitude or level of development of the preference. Meisgeier and Murphy (1987) add, "Neither are scores used to reference any norms; the [Murphy-Meisgeier] is not a normative instrument" (p. 9). Although not normed, substantial samples of individuals from 2nd to 12th grade were used to develop the instrument, with special emphasis placed on the data obtained from 1,506 individuals from grades 2 to 8. The earlier forms of the Murphy-Meisgeier (A, B, C) represent intermediate versions of the instrument that were developed in the course of constructing Form D; Form A was developed from an initial sample of 982 children in grades 3 to 5, and Forms B, C, and D were developed from the subsequent sample of 1,506 individuals.

The Murphy-Meisgeier manual provides split-half and test-retest reliability coefficients for the four indices. Based on a total sample of 720 males and 645 females between grades 2 and 8, the following split-half coefficients were obtained (the value for males is listed first, followed by the value for females): Extraversion-Introversion (.57/.63), Sensing-Intuition (.65/.68), Thinking-Feeling (.59/.63), and Judgment-Perception (.64/.61) (Meisgeier & Murphy, 1987, Table 10.2). The coefficients for the separate grade levels are similar to these overall values. Median test-retest coefficients for 579 of the 1,506 individuals who were retested after four to five weeks are as follows: E-I (.61), S-N (.69), T-F (.58), and J-P (.68) (Meisgeier & Murphy, 1987, Table 18).

Validity data for the Murphy-Meisgeier are meager, basically a set of small to moderate correlations with the Children's Personality Questionnaire (CPQ) scales and factors, and with the Learning Preferences Inventory scores; generally trivial and nonsignificant correlations with the Learning Pattern Assessment; a series of canonical variates based on canonical correlation analysis of the Murphy-Meisgeier with the CPQ (Meisgeier & Murphy, 1987, Table 24) and two learning preference inventories (Fourqurean et al., 1990); and an attempt to establish content validity by obtaining Likert ratings for each Murphy-Meisgeier item from 21 individuals familiar with the concepts of psychological type.

Kaufman Adolescent and Adult Intelligence Test (KAIT). The KAIT (Kaufman & Kaufman, 1993) is a new intelligence test for ages 11 to 85+ years that provides Fluid, Crystallized, and Composite IQs, each with a mean of 100 and standard deviation of 15, and follows the theoretical model of Horn and Cattell (1966, 1967; Horn, 1989). Tasks were developed from the models of Piaget's (1972) formal operations and Luria's (1973) planning ability in an attempt to include high-level, decision-making, adult-oriented tasks. Visual-motor coordination and visual-motor speed are deemphasized, although speed of problem solving is required for several tasks. A Core Battery of six subtests (three Crystallized, three Fluid) yields the three IQs; an Expanded Battery of 10 subtests also includes alternate Crystallized and Fluid subtests, and two tasks that measure the delayed recall of information learned previously in the examination. For the present study, only the IQs were used as variables.

The KAIT was normed on 2,000 individuals aged 11 to 85+ years, and was stratified on the variables of age, gender, race or ethnic group, geographic region, and socioeconomic status (parental education for ages 11-24 years, self-education for ages 25 and above). Mean split-half reliability coefficients were .95 for Crystallized IQ, .95 for Fluid IQ, and .97 for Composite IQ; for ages 11-14, mean values were .92, .94, and .96, respectively. Mean test-retest reliability coefficients, based on 153 normal individuals aged 11-85+ retested after a one-month interval, were as follows: Crystallized IQ (.94), Fluid IQ (.87), and Composite IQ (.94). Values for ages 11-19 were .94, .85, and .95, respectively. Exploratory and confirmatory factor analysis supported the construct validity of the Crystallized and Fluid Scales and the placement of subtests on each scale for each age group including ages 11-14 years. Correlational analyses with the WISC-R at ages 11-16 (N = 118) and the K-ABC at ages 11-12 (N = 124) indicated that KAIT Composite IQ correlated .82 with Full Scale IQ, .66 with K-ABC Mental Processing Composite, and .82 with K-ABC Achievement.

Procedure

Data for this study were obtained during the nationwide standardization of the KAIT between 1988 and 1991. Qualified examiners who were well trained in the administration and interpretation of individual intelligence tests administered the KAIT. Form D of the Murphy-Meisgeier was self-administered by most standardization subjects aged 11 to 14 years, and a few age 15. All Murphy-Meisgeier record forms were machine scored by Consulting Psychologists Press, publisher of the Murphy-Meisgeier.

Data Analysis

Murphy-Meisgeier scores were reported in two ways--categorically, to indicate the direction of the person's preference on each index (if any), and numerically via the "preference scores."

A multiple analysis of variance (MANOVA) was conducted using the following independent variables: Sex, Race/ethnic group (African-American/Anglo-American/Hispanic), and intelligence level; dependent variables were preference scores on the four Murphy-Meisgeier indices. The total sample was divided into three levels of intelligence: 110-160 (N = 62); 90-109 (N = 141); 40-89 (N = 60). The MANOVA was followed by four univariate ANOVAs, one for each index.

Next, a MANCOVA was conducted using Sex, Race/ethnic group, and Fluid-Crystallized IQ discrepancy on the KAIT as independent variables and the four Murphy-Meisgeier indices as dependent variables; educational attainment was the covariate (years of parents' schooling). The total sample was divided into three Fluid (F)-Crystallized (C) discrepancy categories: F > C (N = 50); F = C (N = 174); and C > F (N = 39). The average Fluid-Crystallized IQ discrepancy required for statistical significance at the .05 level is 11 points for ages 11-14 (Kaufman & Kaufman, 1993), so differences of at least 11 points in favor of Fluid IQ were needed to classify a person as F > C; differences of at least 11 points in favor of Crystallized IQ were needed to classify a person as C > F; and differences of 10 points or less resulted in a classification of F = C. The MANCOVA was followed by four univariate ANCOVAs, one for each index.

Educational attainment was used as a covariate in the second set of analyses, but it was undesirable to use it in the first set because education and intelligence are so closely correlated (Kaufman, 1990, Chapter 6); any control for education in the initial analyses would have compromised interpretation of the relationship of intelligence level to Murphy-Meisgeier preferences.

Results and Discussion

The results of the MANOVA are presented in Table 1, and the results of the MANCOVA are presented in Table 2. The independent variable of Sex was significant in the MANOVA ($F = 7.20$, $p < .001$) and remained significant when parents' education (an estimate of socio-economic status) was covaried in the MANCOVA ($F = 5.95$, $p < .001$). Race did not reach significance at the .05 level in the MANOVA, and even though mean parental education differed substantially for the three Race/ethnic groups (14.2 for Anglo-Americans, 13.0 for African-Americans, 10.9 for Hispanics), Race also was a nonsignificant main effect in the MANCOVA. IQ level was a nonsignificant main effect in the MANOVA; Fluid-Crystallized discrepancy was a nonsignificant main effect in the MANCOVA; and all interactions in both multivariate analyses failed to reach significance at the .05 level.

Follow-up ANOVAs and ANCOVAs were conducted to determine which preference scales yielded significant main effects for Sex. Only Thinking-Feeling yielded a significant result in the ANOVA ($F = 18.8$, $p < .001$) or ANCOVA ($F = 21.5$, $p < .001$).

Table 1

Wilks Lambda and *F* Statistics for Each Main Effect and Interaction in the MANOVA of the Four Indices of the Murphy-Meisgeier Type Indicator for Children

Variable	Wilks Lambda	*F*
Sex	.894	7.20***
Race (African-American/Anglo/ Hispanic)	.964	1.12
IQ Level (40-89, 90-109, 110-160)	.954	1.44
Sex X Race	.953	1.46
Sex X IQ	.954	1.46
Race X IQ	.964	0.56
Sex X Race X IQ	.901	1.60

$*p < .05$ $**p < .01$ $***p < .001$

Table 2

Wilks Lambda and *F* Statistics for Each Main Effect and Interaction in the MANCOVA of the Four Indices of the Murphy-Meisgeier Type Indicator for Children

Variable	Wilks Lambda	*F*
Sex	.910	5.95***
Race (African-American/Anglo/ Hispanic)	.942	1.82
Fluid (F)-Crystallized (C) Discrepancy	.970	0.93
Sex X Race	.979	0.68
Sex X F-C	.978	0.68
Race X F-C	.942	0.91
Sex X Race X F-C	.952	1.01

$*p < .05$ $**p < .01$ $***p < .001$

Note. Fluid (F)-Crystallized (C) Discrepancy equals:
F > C: Fluid IQ significantly (11+ pts.) higher than Crystallized IQ ($p < .05$)
F = C: Fluid not significantly different from Crystallized IQ--less than 11 points difference in either direction.
C > F: Crystallized IQ significantly (11+ pts.) higher than Fluid IQ ($p < .05$)

Mean preference scores on the four indexes are presented for various subgroups in Table 3 to help clarify the significant main effect for Sex and to provide data for the main effects that failed to reach significance.

Although preference scores on the four indices are desirable for conducting empirical research on the Murphy-Meisgeier, the person's ratings are usually reported categorically, in terms of the poles that best typify his or her responses. Table 4 shows the percentage of individuals in the sample that were categorized at each pole of the four indices, by sex, race, IQ level, and Fluid-Crystallized discrepancy.

Sex Differences

The significant Sex main effect reflects the fact that females, more so than males, have a decided preference for Feeling rather than Thinking, i.e., they tend to base their judgments on personal values instead of on impersonal analysis and logic. This sex difference is well-documented in the Myers-Briggs literature (Myers & McCaulley, 1985), and was observed in the Myers-Briggs/KAIT study at ages 14 to 94 years

(Kaufman et al., 1992, November). The main difference in the present finding is that both males and females demonstrated a decided preference for Feeling over Thinking (83/7 for females, 53/20 for males). With the Myers-Briggs, males typically show a preference for Thinking. In the Kaufman et al. (1992, November) study, for example, 69% of males were classified as Thinking, and 31% as Feeling.

However, the present finding for 11-15 year-olds does accord well with data reported in the Murphy-Meisgeier manual (Meisgeier & Murphy, 1987, Table 12) for 820 males and 679 females in grades 2-8: The Feeling-Thinking ratios were 81/8 for females and 51/23 for males. It is unclear whether the tendency for males to display a Feeling preference on the Murphy-Meisgeier but a Thinking preference on the Myers-Briggs is a developmental difference pertaining to differences in boys versus men, or a difference in the nature of the two instruments used to study Jungian type. The Murphy-Meisgeier manual fails to present any data for samples of young adolescents who were administered both type indicators; this lack is a serious one, because it impairs the comparison of the present findings on all indices, and for all variables, with data obtained with the same variables on the Myers-Briggs.

Table 3

Means and Standard Deviations of Preference Scores on the Four Indices of the Murphy-Meisgeier Type Indicator for Children, by Sex, Race, KAIT Composite IQ, and KAIT Fluid-Crystallized Discrepancy

		Extravert-E Introvert-I		Sensing-S Intuition-N		Thinking-T Feeling-F		Judging-J Perceiving-P	
Variable/Group	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Sex									
Female	122	46.3	6.5	65.7	8.8	72.6	6.6	69.8	9.1
Male	141	48.8	6.4	65.6	8.2	66.6	6.6	70.1	8.8
Race									
Anglo-American	192	48.2	7.0	66.3	8.6	69.7	7.4	71.1	8.7
African-American	38	46.4	4.5	63.0	8.1	66.7	7.4	65.9	9.1
Hispanic	33	45.9	5.4	64.6	7.6	70.7	5.5	67.9	8.3
IQ Level									
110-160	62	47.5	6.6	68.7	9.0	70.6	7.4	73.8	8.2
90-109	141	47.9	6.9	65.2	8.1	69.1	7.3	69.9	8.7
40-89	60	47.3	5.5	63.4	7.9	68.9	6.9	66.2	8.6
F-C Discrepancy									
F > C	50	46.8	5.7	66.9	7.6	70.1	7.5	70.7	7.6
F = C	174	47.6	6.3	65.1	8.6	69.3	7.1	69.5	9.0
C > F	39	48.8	8.4	66.4	8.7	68.8	7.8	71.0	10.0
Total	263	47.7	6.5	65.6	8.5	69.4	7.2	70.0	8.9
(U-Band)		(47.7-52.3)		(64.4-69.6)		(61.6-66.4)		(63.9-68.1)	
Midpoint		50		67		64		66	

Note. F = Fluid. C = Crystallized. U-Band = Undetermined band. Scores below the lower range for each U-Band denote preferences for the pole listed *first* for each index (Extravert, Sensing, Thinking, Judging); scores above the upper range for each U-Band denote preferences for the pole listed *second* for each pair (Introvert, Intuition, Feeling, Perceiving).

Table 4
Percentage of Subjects Classified at Each Pole of the Four Indices of the Murphy-Meisgeier Type Indicator for Children, by Sex, Race, KAIT Composite IQ, and KAIT Fluid-Crystallized Discrepancy

Variable		Extravert-E Introvert-I		Sensing-S Intuition-N		Thinking-T Feeling-F		Judging-J Perceiving-P	
Group	N	% E	% I	% S	% N	% T	% F	% J	% P
Sex									
Female	122	59.0	17.2	45.1	35.2	6.6	82.8	24.6	57.4
Male	141	45.4	28.4	43.3	27.0	19.9	53.2	22.7	58.9
Race									
Anglo-American	192	48.4	28.1	40.1	34.4	15.1	69.8	18.8	63.0
African-American	38	60.5	7.9	60.5	21.0	15.8	47.4	42.1	42.1
Hispanic	33	60.6	12.1	48.5	21.2	3.0	72.7	30.3	48.5
IQ Level									
110-160	62	51.6	22.6	29.0	48.4	9.7	71.0	11.3	75.8
90-109	60	49.6	24.8	47.5	25.5	16.3	66.0	22.7	56.7
40-89	141	56.7	20.0	51.7	25.0	11.7	65.0	38.3	43.3
F-C Discrepancy									
F > C	50	54.0	18.0	34.0	38.0	10.0	66.0	14.0	60.0
F = C	174	50.0	23.0	47.1	27.6	13.8	66.7	26.4	58.0
C > F	39	56.4	30.8	43.6	35.9	18.0	69.2	23.1	56.4
Total	263	51.7	23.2	44.1	30.8	13.7	66.9	23.6	58.2

Note. F = Fluid. C = Crystallized. The percents of individuals classified at each pole of a given index do not total 100 because some individuals are assigned an Undetermined classification. For example, 59.0% of females were classified as extraverts and 17.2% as introverts, a total of 76.2%; therefore, 23.8% were classified as having an undetermined preference.

Race Differences

The variable of Race was nonsignificant in the multivariate analyses. With the Myers-Briggs, Race was a significant main effect for Thinking-Feeling in an analysis of African-Americans and Anglo-Americans (Kaufman et al. 1992, November) and in an analysis of African-Americans, Anglo-Americans, and Hispanics (Kaufman et al., 1993). African-Americans tended to prefer a Thinking style (71% Thinking/29% Feeling) whereas both Anglo-Americans (48/52) and Hispanics (43/57) were divided about equally on this dimension. The Thinking-greater-than-Feeling result for African-Americans also was evidenced in a study of the Myers-Briggs with 134 African-American high school students from North Carolina enrolled in five science classes (Melear & Pitchford, 1991). For the Murphy-Meisgeier, all three Race/ethnic groups demonstrated a strong preference for the Feeling dimension (see Table 4). The result is not statistically significant, but note from Table 4 that African-Americans had a smaller percent classified as Feeling (47) than either Anglo-Americans (70) or Hispanics (73), as was found in the Kaufman et al. (1992, November) Myers-Briggs study.

Levy, Murphy, and Carlson (1972) found dramatic ethnic differences when they compared 758 African-American college students to 3,916 Anglo-American college students. African-American college students had a predominance of Sensing and Judging types, relative to their Anglo-American counterparts; this finding held both for males and females. Again, the data in Table 4 for Race represent nonsignificant findings, but the trend is consistent with the Levy et al. (1972) results: Among African-Americans, 60% were Sensing types compared to 40% of Anglo-Americans; and 42% of African-Americans were Judging types compared to 19% of Anglo-Americans.

KAIT IQ Level and Fluid-Crystallized Discrepancy

The nonsignificant IQ level main effect for the Murphy-Meisgeier differs from the significant main effect that emerged in the Myers-Briggs/KAIT study (Kaufman et al., 1992, November) for Sensing-Intuition. In that study, people relied much more on Sensing than Intuition if they had average intelligence (72%/28%) or below average intelligence (86%/14%), whereas people with high intelligence relied upon Sensing (51%) and Intuition (49%) about equally. Individuals with average or below average IQs tend to perceive the environment by reporting observable facts or happenings through one or more of the five senses; they were much less likely than intelligent people to report possibilities and relationships. Similarly, much previous research indicated that the Sensing-Intuition dimension, much more so than the other three Myers-Briggs dimensions, was related to the cognitive ability of adolescents and young adults on group-administered tests such as the Scholastic Aptitude Test and National Teacher's Examination, and on other measures such as grade point average (Myers & McCaulley, 1985; Pratt, Uhl, Roberts, & DeLucia, 1981; Schurr, Ruble, & Henriksen, 1988).

The investigation of the Murphy-Meisgeier does not confirm these previous findings because of the nonsignificant main effect for IQ level in the multivariate analyses. However, a trend evident in Table 4 conforms to the persistent Myers-Briggs finding of greater dependency on an Intuitive perception of the environment for bright, relative to less bright, individuals. Among individuals with IQs of 110 and above, 48% were classified as Intuitive and 29% as Sensing types; these percents were approximately reversed for the other two IQ levels. The association of high IQ with the Intuitive type is sensible in view of the descriptions of the two types. Intuition is described with statements such as "Focuses on concepts," "Enjoys learning new skills," and "Looks for new ways of doing things"; Sensing is described with statements such as "Likes things definite and measurable" and "Trusts customary ways of doing things" (Murphy & Meisgeier, 1987, Table 1).

The variable of Fluid-Crystallized discrepancy was not significant in the multivariate analysis, and an examination of Tables 3 and 4 reveals that the percents classified at each pole of the typologies were quite similar for the three discrepancy categories. These results are quite consistent with the Myers-Briggs findings (Kaufman et al., 1992, November). They suggest that for the entire adolescent and adult age range covered by the KAIT, the Horn-Cattell constructs of fluid and crystallized intelligence--as measured by the KAIT scales of the same name--are apparently independent of the Jungian constructs, as measured by the Murphy-Meisgeier and Myers-Briggs.

Age Differences

The biggest age difference between the present study and the previous Myers-Briggs study (Kaufman et al., 1992, November) concerns the Thinking-Feeling dimension. About 67% of 11-15 year-olds in this study demonstrated a Feeling preference, and only about 14% had a Thinking preference. To make these data more comparable to Myers-Briggs data, the group of Undetermined individuals should be eliminated from consideration. When considering only those people who received a classification, then 17% were categorized as Thinking and 83% as Feeling. The ratio of 83:17 Feeling-to-Thinking is far different from the ratio of 52:48 for the youngest age group in the Kaufman et al. (1992, November) study (14-19 years) and from the adult age groups in that study (50:50 for 20-29 and 30-49, and 54:46 for 50-94). This difference may be developmental, but as mentioned previously, no equating study between the Myers-Briggs and Murphy-Meisgeier has been made available; consequently, the differences noted may be instrument-related and not age-related. In addition to being composed of different items, the two type indicators differ in their method of assigning weights to items and in the decision of whether or not to force people to be categorized at one pole or the other.

Kaufman et al. (1992, November) identified one significant main effect for Age in the Myers-Briggs study--Judging-Perceiving, which resulted from an apparent age trend: People tended to become more judging, and less perceiving, with increasing age. Individuals aged 14-19 relied more on perception than judgment, dealing with the outer world more by sensing or intuition than by thinking or feeling. In contrast, individuals aged 30 and above relied more on judgment (what they think or feel) than on perception (what they sense or intuit). If the Undetermined category is eliminated, and the Murphy-Meisgeier classifications are recomputed for Judging-Perceiving, then the obtained percentages continue the significant age relationship that was found in the Myers-Briggs, as shown below (data for the Myers-Briggs are from Kaufman et al., 1992, November):

Test	Percent Judging	Percent Perceiving	Age Group
Murphy-Meisgeier	28.9	71.1	11-15
Myers-Briggs	39.4	60.6	14-19
Myers-Briggs	50.0	50.0	20-29
Myers-Briggs	62.3	37.7	30-49
Myers-Briggs	68.3	31.7	50-94

Age-difference data for eight age groups between 15-17 years and 60+ years from the Myers-Briggs data bank for Form F (N>50,000) and Form G (N > 30,000) (Myers & McCaulley, 1985, Appendix C) are quite consistent with the age relationships depicted above.

The Murphy-Meisgeier manual (Meisgeier & Murphy, 1987) discusses the importance of the development of type, and states, "According to current belief, the child's dominant type emerges sometime between the ages of 6 and 14" (p. 7). The authors also gathered data on a substantial sample of over 1,500 children in grades 2 to 12, and presented reliability estimates for grades 2 through 8. Yet they did not present any data for separate age or grade levels on the four indices that would allow the examination of possible developmental changes in typology. Such data are needed. The present study afforded comparisons between 11-15 year-olds and several adolescent and adult groups on the Myers-Briggs. But the present age group was too narrow to permit meaningful evaluation of age trends within the Murphy-Meisgeier. Additional research on the Murphy-Meisgeier, with a wide age range of children and adolescents, is essential for proper interpretation of the instrument in school settings.

Conclusions

The results of this study indicate that knowledge of an individual's sex is an important aspect of interpreting the Thinking-Feeling index on the Murphy-Meisgeier, a finding that has been previously known for this instrument, and is well-known for the Myers-Briggs. However, knowledge of an individual's KAIT Composite IQ, race or ethnic group, and discrepancy between fluid and crystallized abilities will not modify an examiner's interpretation of the inventory. That is to say, individuals who administer the Murphy-Meisgeier to preadolescents and young adolescents are able to interpret the profiles in much the same way regardless of the person's level of intelligence, pattern of displaying that intelligence, or racial/ethnic background. That degree of generalizability was not found to hold true for adolescents and adults on the Myers-Briggs, since significant relationships were obtained with Race/ethnic group (Kaufman et al., 1993; Kaufman et al., 1992, November) and with IQ level (Kaufman et al., 1992, November), as well as with Sex and Age (Kaufman et al., 1992, November). Whether the present findings with the Murphy-Meisgeier apply to children below age 11 cannot be answered by the present data.

Implications

The results of this study have implications for classroom teachers, counselors, and other school leadership personnel. A recent report commissioned by the American Association of University Women (AAUW) Educational Foundation titled *How Schools Shortchange Girls* found that "despite a narrowing of the 'gender gap' in verbal and mathematical performance, girls are not doing as well as boys in our nation's schools" (Wellesley College Center for Research on Women, 1992, p. 16). In contrast to the claims of some researchers (e.g., Jacklin, 1989) who suggest that gender differences are inconsequential and that gender research should cease, the report indicated that gender differences in science achievement are not decreasing and may be increasing. Even more relevant to the present study was a finding that a decidedly higher percentage of males had career plans for engineering and the physical sciences while a higher percentage of females planned on a career in the social sciences (Wellesley College Center for Research on

Women, 1992). The report also presented evidence that there is no "math gene" to account for these differences, a finding supported by Jacklin (1989).

It must be recognized that gender differences may impact on learning. As found in this study, girls and boys between the ages of 11 and 15 (fifth through ninth grades) have real differences in how they make judgments. The boys' approach may be more conducive to success in math and science while the girls' approach may be more conducive to success in the arts and social sciences. If the results of the present study are compared to those of adults (Kaufman et al., 1992, November), the proportion of males who favor the Feeling perspective is greatly reduced. A recommendation of the AAUW report is that "testing and assessment must serve as stepping stones not stop signs" (The Wellesley College Center for Research on Women, 1992, p. 87). With this recommendation in mind, how can the results of this study help educators? While the differences reported in this study are statistically significant, there is a wide variability in each gender. Therefore, consider administering the Murphy-Meisgeier or Myers-Briggs to anyone who might have difficulty in math/science or the arts/social sciences to infer whether dimensions of personality are influencing educational outcomes. This approach would provide teachers and other educators with the opportunity to provide appropriate interventions.

While it would be difficult to suggest specific intervention strategies in these situations, a number of ideas may help students with a preference for Feeling or Thinking to understand and improve their functioning in the other direction on the continuum. Teachers should offer equal opportunities with Feeling versus Thinking reactions in the classroom. Studies (e.g., Van, 1992) have shown that individuals with a preference for Thinking "meet with academic success due, in large part, to their propensity for systematic analysis and their ability to make decisions based on pertinent facts. . . . Feeling types, however, want topics they can care about and assignments that they believe have value" (p. 23). Most questioning is geared to a "one correct answer" response rather than open-ended options, inviting responses mostly on the Thinking level. One strategy would be to have more open-ended questions, call on individual students one at a time, and provide adequate time for response. More opportunities could be offered for multisensory options in evaluation and class assignments (e.g., acting the scene in history or drawing a mural versus talking or short answer tests). The current trend toward cooperative learning gives students the chance to work together; girls who show a preference for Feeling can add that perspective to group discussions rather than relying on teacher lecture with student response (often geared to the Thinking perspective).

The traditional approaches to math and science reinforce the Thinking perspective. Alternative teaching methods and philosophies (e.g., constructivist view of knowledge) encourage teachers and students to consider the Feeling perspective. Research (Bailey, 1993) shows that teachers (both male and female) give preference to males in the classroom and this is even more pronounced in science classes. In recognizing the Feeling perspective, teachers should allow time and make efforts to give equal opportunities to all students.

Parents should not be left out of the picture. Parents as well as teachers should model both Thinking and Feeling strategies without regard to the sex of the child. Fathers should try to model Feeling processes and mothers Thinking. Male teachers should allow opportunities for students to see them "talking through" the Feeling process. In other words, both teachers and parents can avoid influencing only one Thinking/Feeling preference in children of each sex. Findings such as those from the present study should be used in conjunction with appropriate intervention strategies to reduce gender inequities. Understanding that the root of many of these inequities may be a personality trait could help to reduce this gap.

References

Allen, T. (1989). Minnoka schools lead the way in exploring the idea. *School Administrator, 46* (9), 8-11.

Bailey, S. M. (1993). The current status of gender equity research in American schools. *Educational Psychologist, 28*(4), 321-339.

Briggs, K. C., & Myers, I. B. (1983). *Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Carlson, J. G. (1985). Recent assessments of the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 49,* 356-365.

Carlson, J. G. (1989). Affirmative: In support of researching the Myers-Briggs Type Indicator. *Journal of Counseling and Development, 67,* 484-486.

DeVito, A. J. (1985). Review of Myers-Briggs Type Indicator. In J. V. Mitchell, Jr. (Ed.), *The ninth mental measurements yearbook* (pp. 1030-1032). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.

Dilley, J. (1987). Applications of Jungian type theory to counselor education. *Counselor Education and Supervision, 27*, 44-52.

Fourqurean, J. M., Meisgeier, C., & Swank, P. (1990). The link between learning style and Jungian psychological type: A finding of two bipolar preference dimensions. *Journal of Experimental Education, 58*, 225-237.

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research.* New York: W. H. Freeman.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253-270.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129.

Jacklin, C. N. (1989). Female and male: Issues of gender. *American Psychologist, 44*, 127-133.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston. MA: Allyn and Bacon.

Kaufman, A. S., & Kaufman, N. L. (1993). *Manual for the Kaufman Adolescent and Adult Intelligence Test (KAIT).* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Kaufman, N. L., & McLean, J. E. (1993). Profiles of Hispanic adolescents and adults on the Myers-Briggs Type Indicator. *Perceptual and Motor Skills, 76*, 628-630.

Kaufman, A. S., McLean, J. E., & Underwood, S. S. (1992, November). *The Myers-Briggs Type Indicator: Relationships to IQ level and fluid-crystallized discrepancy on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at ages 14 to 94 years.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Levy, N., Murphy, C., Jr., & Carlson, R. (1972). Personality types among Negro college students. *Educational and Psychological Measurement, 32*, 641-653.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology.* New York: Basic Books.

Lynch, A. G. (1985). The Myers-Briggs Type Indicator: A tool for appreciating employee and client diversity. *Journal of Employment Counseling, 22*, 104-109.

Meisgeier, C., & Murphy, E. (1987). *Murphy-Meisgeier Type Indicator for Children manual.* Palo Alto, CA: Consulting Psychologists Press.

Melear, C. T., & Pitchford, F. (1991, July). African-American science student learning style. In *Proceedings of the International Conference of the Association for Psychological Type* (9th, Richmond, VA).

Murphy, E. A. (1986). Estimation of reliability and validity for the Murphy-Meisgeier Type Indicator for Children. *Dissertation Abstracts International*, 86-96-64.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15*, 1-12.

Pratt, L. K., Uhl, N. P., Roberts, A. R., & DeLucia, S. (1981, May). *The relationship of the Myers-Briggs Type Indicator to scores on the National Teacher's Examination.* Paper presented at the Twenty-first Annual Forum of the Association for Institutional Research, Minneapolis, MN. (ERIC Document Reproduction Service No. ED 205 128)

Schurr, K. T., Ruble, V. E., & Henriksen, L. W. (1988). Relationships of Myers-Briggs Type Indicator personality characteristics and self-reported academic problems and skill ratings with Scholastic Aptitude Test (SAT) scores. *Educational and Psychological Measurement, 48*, 187-196.

Van, B. (1992). The MBTI: Implications for retention. *Journal of Developmental Education, 16*(1), 20-25.

Wellesley College Center for Research on Women. (1992). *How schools shortchange girls.* Washington, DC: American Association of University Women Educational Foundation.

Wiggins, J. S. (1989). Review of the Myers-Briggs Type Indicator. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurements yearbook.* (pp. 537-538). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska Press.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 49-58

A Comparison of Two Procedures, the Mahalanobis Distance And the Andrews-Pregibon Statistic, for Identifying Multivariate Outliers

Michele Glankler Jarrell

The purpose of this study was to compare two procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, for identifying multivariate outliers under varying conditions of extremeness and dimension. The null hypotheses were whether there would be a significant difference between procedures, between degrees of extremeness, and among dimensions in identifying outliers. From a three-dimensional multivariate normal population, 1,100 samples were computer-generated. Outliers were induced according to varying combinations of extremeness and dimension, producing 6,600 samples. Each procedure was run; data on outliers were compiled; and ANOVA was run with false outliers and total outliers identified as dependent variables. The procedures, degrees of extremeness, and dimensions were all statistically significant. The results were analyzed for practical significance using eta-square. Procedure accounted for less than 1% of the variability. There was a significant difference between degrees of extremeness and among dimensions. The conclusion is that choice of procedure is not critical. Both procedures identified valid data points as outliers. Due to these results with false outliers as the dependent variable, it is recommended that the researcher investigate the results of any outlier identification procedure before determining the fate of suspect observations.

Introduction

Examination of outliers is an essential part of any analysis, univariate or multivariate. The results of many classical statistical procedures can be distorted by the occurrence of outliers; "estimators that are optimal under a Gaussian [normal] assumption are very vulnerable to the effects of outliers" (Wainer, 1976, p. 285); therefore the identification and possible removal or accommodation of outliers are important considerations. Often the outliers may be the point of interest, as in identifying exceptional schools in a system or exceptional teachers in a field. An outlier may simply be the result of an error in observation or data entry; in this case, identification would permit the researcher to make appropriate corrections. Gnanadesikan (1977) stated that "the consequences of having defective responses are intrinsically more complex in a multivariate sample" (p. 271) than in a univariate sample.

"Outliers occur very frequently in real data" according to Rousseeuw and Leroy (1987, p. vii). Huber (1977) stated that having 5% to 10% "wrong values" (p. 3) is the norm. Thus, procedures for finding these outlying values and for deciding how to handle them are essential. Whatever the cause of the outliers, we must identify them in order to decide how best to deal with them in the statistical analysis. Wood (1983) saw three possibilities for dealing with outliers after they have been identified: remove them from the data set, keep them in order to extend our range of knowledge, or modify the model to accommodate them. Chatterjee and Hadi (1988) recommended the outliers be examined for "accuracy . . ., relevancy . . ., or special significance" (p. 182).

Outliers require consideration; as several authors have indicated, they are an unavoidable problem (Barnett & Lewis, 1978; Douzenis & Rakow, 1987; Huber, 1977). If one deals with data, one must be able to identify outliers and to decide how to treat them. Data are often scanned into a computer file for analysis thereby becoming "invisible" (Gentleman & Wilk, 1975, p. 387), or data are in such large quantities that they are impossible to inspect visually. Procedures to identify possible outliers in large multivariate data sets are a necessity.

Definitions of Outlier

Many of the researchers who have dealt with the problem of outliers have based their work on a subjective definition of an outlier. Outliers have been seen as values that are "dubious in the eyes of the analyst" (Dixon, 1950, p. 488) or that appear "to deviate markedly from other members of the sample" (Grubbs, 1969, p. 1); and Elashoff and Elashoff (1970) stated outliers are observations that are "extreme in some sense" (p. 4). Many other researchers have used the same basic definition

---

Michele G. Jarrell is an Associate Research Educational Psychologist with the Evaluation and Assessment Laboratory in the College of Education at The University of Alabama. Please address correspondence regarding the paper to Michele G. Jarrell, EAL, The University of ALabama, Box 870231, Tuscaloosa, AL 35487-0231
(Internet: MJARRELL@CCMAIL.BAMANET.UA.EDU)

(Barnett, 1978; Barnett & Lewis, 1978; Pascale & Lovas, 1976; Rasmussen, 1988; Robertson, 1987).

Guttman (1973) saw an outlier as a spurious observation that did not come from a $N(\mu,\sigma^2)$ population. Gentleman and Wilk (1975) pointed out that an outlier could be an outlier only "relative to some prespecified model or theory . . ." (p. 389). Hawkins (1980) defined multivariate outliers as "values with high probabilities of occurring where the probability density of the true distribution is low, remote from the main body of data" (p. 104).

Many definitions are based on the type of analysis; for instance, in a regression analysis, an outlier is a value which deviates from the regression line (Douzenis & Rakow, 1987) or one with a high residual (Chatterjee & Hadi, 1988). Rousseeuw and van Zomeren (1990) see outliers as "observations that deviate from the model suggested by the majority of the point cloud " (p. 651).

Causes of Outliers

Among the most commonly cited reasons for the occurrence of outliers in data are errors in collecting, recording, coding, or entering data, and deviations from the experimental design (Chatterjee & Hadi, 1988; Douzenis & Rakow, 1987; Portnoy, 1988; Seber, 1984); Barnett and Lewis (1978) referred to these as human error and ignorance. These are the outliers that require identification in order to be corrected or rejected. Some outliers occur due to violations of the assumptions; they may indicate the model is not an appropriate one for the data, and they will affect the inferences drawn from the procedures used. Outliers may be due to the "variability inherent in the data" (Grubbs, 1969, p. 1) as with data from a "heavy tailed distribution such as Student's *t*" (Hawkins, 1980, p. 1); in this case, the "outliers" are actually valid data points and should not be deleted. Data may actually be from two populations with different distributions, in which case the outliers would be observations not from the basic distribution. These outliers should be rejected or given small weights (Hawkins, 1980).

Identification of Outliers

Identification of outliers is critical because "many of the standard multivariate methods are derived under the assumption of normality and the presence of outliers will strongly affect inferences made from normal-based procedures" (Schwager & Margolin, 1982, p. 943).

There are several procedures for the identification of multivariate outliers, each with its adherents and detractors (Comrey, 1985; Grubbs, 1969). The question addressed in this study will be how two procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, compare in detecting outliers in a multivariate normal population under varying conditions. These two procedures were selected for two reasons: Since one procedure is based on distance and the other on volume, they should identify the same or similar observations as outliers; and both of the procedures are easily programmed on the computer.

Purpose of the Study

Presently, many authors suggest that researchers use one or more procedures for identifying outliers before performing their statistical analyses (Gnanadesikan & Kettenring, 1972; Krzanowski, 1988; Stevens, 1984). Outliers and influential data can be the most important data in an analysis and are deserving of special attention according to Gray (1989). The points classified as outliers may differ dramatically according to which identification procedure is used.

This study used a factorial design to compare the results of two procedures for identifying multivariate outliers under varying conditions. Results were analyzed for the total number of outliers identified and for the number of false outliers identified. Simulated data were generated by computer and were limited to three dimensions (e.g., three uncorrelated variables). Using known population parameters, 1,100 samples of size 150 were generated. A sample size of 150 was selected to approximate the sample size of many studies in the behavioral sciences. Outliers were induced by replacing randomly selected data points in each sample with plus or minus the value of three or six standard deviations from the mean. The samples had outliers induced into one dimension; then the samples had outliers induced into two dimensions; and finally the samples had outliers induced into all three dimensions. Both procedures were used on all samples.

Definition of Terms

Outlier. An extreme observation, either high or low, which does not conform to the distribution of the majority of observations in the sample. An observation from a distribution with different parameters than the majority of the observations. For the purpose of this study, an outlier will be defined as a data point plus or minus either three or six standard deviations.

False outlier. An observation occurring naturally in the population but identified by the procedure as an outlier.

Hypotheses

The study tested the following null hypotheses using as dependent variables both the total number of outliers detected and the number of false outliers detected. The total number of outliers included both the induced outliers and the false outliers.

Ho_1: There will be no significant differences at the .05 level between the two procedures in detecting outliers.

Ho_2: There will be no significant differences at the .05 level between the "Three Standard Deviation" and the "Six Standard Deviation" outlier groups in the number of outliers detected.

Ho_3: There will be no significant differences in the number of outliers detected at the .05 level among the groups with outliers in one, two, and three dimensions.

As the performance of the two procedures was the main interest of this study, there were no hypotheses about the interaction effects.

Method

Type of Study

This was an empirical study that used computer generated data. Shapiro, Wilk, and Chen (1968) found that "empirical sampling using a high speed computer can provide a very useful general guide on sensitivity properties even with a few Monte Carlo runs (e.g., 100, 200, or 500)" (p. 1371). Two methods of detecting multivariate outliers, the Mahalanobis distance and the Andrews-Pregibon statistic, were compared under varying conditions for a set of population parameters. The population was a large three dimensional multivariate normal data set with known means and variances.

Generation of Data

The data set was generated using a FORTRAN program on the IBM 3090/400E computer at The University of Alabama Seebeck Computer Center. Morris (1975) recommended using his FORTRAN program to generate data for use in Monte Carlo studies; the program creates a population with the desired centroid and covariance matrix. Using the population parameters, 1,100 random samples were generated. Setting σ at the maximum value of .5 and solving the formula

$$\text{bound of error (BOE)} = 2\frac{\sigma}{\sqrt{n}}$$

for n, a BOE of .05 required an n of 400, and a BOE of .04 required an n of 625. Using 1,100 samples gave a bound on the error of between .03 and .301 or 3 and 3.1 percentage points. The 1,100 random samples from the population were generated on the computer after setting the population parameters in the FORTRAN program. In order to generate data that were consistent with the type of data a behavioral scientist would see, the population parameters used in the study were those of the standard score distribution, that is, multivariate normal (0,1). These population parameters provided for simplicity of interpretation.

The sample sizes were 150. In order to induce outliers in the samples, the following method was used. The standard deviations obtained from the selected population parameters were multiplied by three and by six to obtain constants that replaced, as a positive or negative value, eight observations in the original samples. According to Huber (1977), 5% to 10% of values might be outliers, therefore 8 outliers (5%) were induced in each sample of 150. Since the samples were randomly generated, the first eight observations in each sample were replaced. For the samples with outliers in one dimension, the appropriate constant (3σ or 6σ of the first variable) replaced the first variable of four observations, and the negative of the appropriate constant replaced the first variable of four observations. For the samples with outliers in two dimensions, the appropriate constants (3σ or 6σ of the first and second variables) replaced the first two variables of two observations; the negative of the constant replaced the first two variables of two observations; the constant replaced the first, and the negative constant replaced the second variable of two observations; and the negative constant replaced the first, and the constant replaced the second variable of two observations. For samples with outliers in three dimensions, each of eight multivariate outliers was induced according to a different pattern. The appropriate constants (3σ or 6σ of the three variables) replaced all three variables in one observation, and the negative constants replaced all three variables in one observation. In one observation, the constants replaced the first two variables, and the negative constant replaced the third; in one observation, the constant replaced the first variable, and the negative constants replaced the last two; in one observation, the negative constant replaced the first variable, and the constants replaced the last two; in one observation, the negative constants replaced the first two variables, and the constant replaced the third. In one observation, the constants replaced the first and third variables, and the negative constants replaced the second; in one observation, the negative constants replaced the first and third variables, and the constant replaced the second. Table 1 illustrates the pattern for inducing outliers.

Each sample was manipulated in terms of the number of dimensions and the extremeness of the outliers; therefore, there were actually six outlier samples produced from each original sample.

Table 1
Clustering of Outliers According to Dimensions

Dimension		
One	Two	Three
+	+ +	+ + +
+	+ +	- - -
+	- -	+ + -
+	- -	- - +
-	+ -	+ - -
-	+ -	- + +
-	- +	+ - +
-	- +	- + -

where + is $x_{(i)} = +3\sigma$ or $x_{(i)} = +6\sigma$ and
where - is $x_{(i)} = -3\sigma$ or $x_{(i)} = -6\sigma$

The data were generated on an IBM 3090\400E mainframe computer using a REXX executable file (International Business Machines, 1988a; International Business Machines, 1988b) to run several SAS programs (SAS Institute Inc., 1990a; SAS Institute Inc., 1990b) and a FORTRAN program. Eleven hundred samples of n = 150 and dimension = 3 were generated from a multivariate normal (0,1) population using a FORTRAN program developed by Morris (1975). The Mahalanobis distance (Stevens, 1984) and the Andrews-Pregibon statistic (Andrews & Pregibon, 1978) were calculated for each observation in each sample. A data file was created with the statistics and the row number from the matrix for each of the two procedures for each sample.

The FORTRAN data listing was transformed into SAS/IML matrix form and written into a data file. The SAS program that calculated the Mahalanobis distance was edited using an executable file that substituted the new data matrix during each run through the programs. When the SAS program was run, it induced the outliers into the data matrix, calculated the Mahalanobis distances, and produced a listing of the distances for the matrix. Another SAS program read the output and appended the data to a data file if the Mahalanobis distance met or exceeded the critical value; therefore, after the 1,100 runs through the REXX executable file, there was one data file containing the Mahalanobis distance and row number for each observation identified as an outlier. The SAS program that ran the Andrews-Pregibon statistic on the matrix was edited using an executable file that substituted the new data matrix during each run through the program. When the SAS program was run, it deleted one observation at a time from the data matrix, calculated the Andrews-Pregibon statistic, and produced a listing of the statistics and the row number from the matrix. Another SAS program read the output and appended the data to a data file if the Andrews-Pregibon statistic was equal to or less than the critical value; therefore, after the 1,100 runs through the REXX executable file, there was one file containing the Andrews-Pregibon ratio and the row number for each observation identified as an outlier.

The two data files containing the Mahalanobis and Andrews-Pregibon outlier data were downloaded to a microcomputer. Two BASIC programs were run. The first program read the Mahalanobis outlier data and produced another data file containing the sample number and the number of false outliers, induced outliers, and total outliers identified under each combination of extremeness and dimensionality. Thus, for each sample there were six lines in the output data file. The second BASIC program did the same thing for the Andrews-Pregibon outlier data. Data lines in the output files were coded "1" for Mahalanobis data and "2" for Andrews-Pregibon data, then the Andrews-Pregibon data were appended to the Mahalanobis data forming a single data file. This file was uploaded to the mainframe and sorted by procedure and sample. The sorted file contained 12 lines of data for each of the 1,100 samples.

Outlier Identification Procedures

The first procedure, the Mahalanobis distance, gives the "distance from the case to the centroid of all cases for the predictor variables" (Stevens, 1984, p. 339). A large distance in relation to the other distances obtained indicates an outlier. This distance also measures an observation's leverage (Chatterjee & Hadi, 1988). The

Mahalanobis distance, D_i^2, is expressed in terms of the covariance matrix S:

$$D_i^2 = (X_i - \bar{X})^T S^{-1} (X_i - \bar{X})$$

In this formula, X_i is the vector of data for case i, and $\bar{X}$ is the vector of means for the predictors. Under the assumption that the predictors came from a multivariate normal population, the critical values of D_i are given by Barnett and Lewis (1978) for alpha = .05 and .01 and for $p = 2$ to 5, where p is the number of dimensions.

For each sample, the induced outliers were added in one dimension for the first degree of extremeness (i.e., three standard deviations), the Mahalanobis distance was calculated, and a vector of the values was printed. The vector of Mahalanobis distances was read, and any value equal to or exceeding the critical value of 7.81473 (Tabachnick & Fidell, 1983, p. 479) was printed with the row number indicating which observation it had identified. The same procedure was followed for each of the dimensions and each degree of extremeness. Therefore, for each of the 1,100 samples, there was a listing of the possible outliers with their corresponding Mahalanobis distance for each of the six possible combinations of dimensionality and extremeness.

The second procedure, the Andrews-Pregibon statistic, is based on the volume of the confidence ellipsoid. Andrews and Pregibon (1978) stated that this procedure identifies observations that are potential outliers and that are influential on the linear model estimates. The Andrews-Pregibon ratio is expressed as:

$$AP = \frac{\det\,(X_{(i)}^T X_{(i)})}{\det(X^T X)}$$

where *det* is the determinant of the matrix that results from multiplying the two matrices, X^T is the transpose of the X matrix, $X^T_{(i)}$ is the transpose of the X matrix with the ith observation deleted, and $X_{(i)}$ is the X matrix with the ith observation deleted.

For each sample the induced outliers were added in one dimension for the first degree of extremeness, i.e., three standard deviations, the Andrews-Pregibon statistic was calculated, and a vector of the values was printed. The vector of Andrews-Pregibon statistics was read, and any value less than or equal to the critical value of 0.9484 (Jarrell, 1991) was printed with the row number indicating which observation it had identified. The same procedure was followed for each of the dimensions and each degree of extremeness. Therefore, for each of the 1,100 samples, there was a listing of the possible outliers with their corresponding Andrews-Pregibon statistic for each of the six possible combinations of dimensionality and extremeness.

Data Analysis

Analysis of variance was run using the total number of outliers identified in each sample by the two procedures and the number of false outliers identified in each sample by the two procedures as the dependent variables. The independent variables were the outlier identification procedure, the extremeness of the outliers, and the number of dimensions in which the outliers occurred. The data were entered in a factorial design with procedure crossed with extremeness of outliers and with number of dimensions in which the outliers occurred. The data were analyzed according to the recommendations of Looney and Stanley (1989) and Barcikowski and Robey (1984).

The research design layout is shown in Table 2.

Table 2
Research Design Matrix

Dimensions	One				Two				Three			
Extremeness	3σ		6σ		3σ		6σ		3σ		6σ	
Procedure	1	2	1	2	1	2	1	2	1	2	1	2
Sample$_1$												
.												
.												
.												
Sample$_{1,100}$												

Tukey's Honestly Significant Difference (HSD) Procedure (SAS Institute, 1990c) was run on the main effect of dimensionality in order to compute the minimum significant difference.

Results

The analysis of variance procedure was run on the data generated by the two outlier identification procedures to partition the variance accounted for by the design. As expected, the large sample size (n = 1,100) resulted in all the statistical tests being significant at the $p < 0.0001$ level; therefore, interpretation of the study is based on practical significance as indexed by eta-square rather than on statistical significance. An eta-square greater than or equal to .10 (10%) is considered to be significant. In terms of practical significance, only the main effects and one two-way interaction were found to be significant. Results were analyzed separately for the two dependent variables, the number of false outliers identified, and the total number of outliers identified.

False Outliers

For the false outliers identified, the overall model and each of the variables were statistically significant at the 0.0001 level. The F-value for each variable is listed in Table 3. The overall model accounted for 84.4% of the total variability; subtracting the variability accounted for by the sample (7.9%) leaves 76.2% accounted for by dimensionality (18.6%), degree of extremeness (47.3%), and the interaction of dimensionality and extremeness (10.3%). The procedure for identifying outliers accounted for less than 1% of the variability, as did the interactions of procedure with dimensionality, procedure with extremeness, and the three-way interaction of procedure, dimensionality, and extremeness.

Table 3
ANOVA Summary Table for False Outliers

Source	df	Sum of Squares	Mean Square	*F* Value	R-Square
Model	1110	41215.99	37.13	58.85*	0.8438
Error	12089	7627.18	0.63		
Corrected Total	13199	48843.16			

Source	df	Sum of Squares	Mean Square	*F* Value	R-Square
Dimensionality	2	9068.94	4534.47	7189.09*	0.1857
Extremeness	1	23108.05	23108.05	3984.48*	0.4731
Ext X Dim	2	5027.77	2513.89	3964.48*	0.1029
Procedure	1	54.61	54.81	86.55*	0.0011
Dim X Proc	2	32.50	16.25	25.75*	0.0007
Ext X Proc	1	34.93	34.93	55.36*	0.0007
Ext X Dim X Proc	2	26.28	13.14	20.83*	0.0005
Sample	1099	3862.91	3.51	5.57*	0.0791

* $p < 0.0001$

Identifying false outliers represents error on the part of the outlier identification procedures. In a practical sense, the choice of procedure was not an issue; although statistically significant, the choice of procedure accounted for only about one-tenth of one percent of the variability. The degree of extremeness of the outliers accounted for most of the variability in the model. The six standard deviations and the three standard deviations degree of extremeness were significantly different at the 0.0001 level and accounted for about 47% of the variability. The dimensionality accounted for a significant amount of the variability. There was a significant difference between the two procedures and between the two degrees of extremeness, and Tukey's HSD at the .05 level showed a minimum significant difference of 0.0397 among the three dimensions. Values obtained for Tukey's HSD can be found in Table 4.

Table 4
Tukey's HSD for False Outliers

Dimensionality	Mean *	
1	2.58409	Minimum significant difference = 0.0397
2	1.61045	
3	0.55432	

* All means are significantly different

An inspection of several of the data sets showed that both outlier identification procedures selected similar numbers of false outliers in each of the six combinations of dimensionality and extremeness and many times selected the same observations as outliers. Table 5 lists the means and standard deviations for the number of false outliers identified by the procedures.

Total Outliers

For the total number of outliers identified, the overall model and each of the variables were statistically significant at the 0.0001 level. The *F*-value for each variable is listed in Table 6. The overall model accounted for 61% of the total variability; subtracting the variability accounted for by the sample (17.5%) leaves 42.7% accounted for by dimensionality (10.7%), degree of extremeness (18.1%), and the interaction of dimensionality and extremeness (13.9%). The procedure accounted for less than 1% of the variability, as did the interactions of procedure with dimensionality, procedure with extremeness, and the three-way interaction of procedure, dimensionality, and extremeness.

Table 5
Means and Standard Deviations for False Outliers

Procedure	Dimension	Extremeness	Mean*	Standard Deviation
Mahalanobis	1	3σ	4.38	1.43
Andrews-Pregibon	1	3σ	4.87	1.52
Mahalanobis	2	3σ	2.94	1.19
Andrews-Pregibon	2	3σ	3.06	1.24
Mahalanobis	3	3σ	1.05	0.90
Andrews-Pregibon	3	3σ	1.13	0.95
Mahalanobis	1	6σ	0.52	0.70
Andrews-Pregibon	1	6σ	0.56	0.73
Mahalanobis	2	6σ	0.20	0.40
Andrews-Pregibon	2	6σ	0.24	0.43
Mahalanobis	3	6σ	0.02	0.14
Andrews-Pregibon	3	6σ	0.02	0.14

* All means are based on *n*'s of 1,100

Table 6
ANOVA Summary Table for Total Outliers

Source	df	Sum of Squares	Mean Square	*F* Value	R-Square
Model	1110	18703.42	16.85	17.16*	0.6118
Error	12089	11868.27	0.9817		
Corrected Total	13199	30571.69			

Source	df	Sum of Squares	Mean Square	*F* Value	R-Square
Dimensionality	2	3262.78	1631.39	1661.73*	0.1067
Extremeness	1	5534.18	5534.18	5637.11*	0.1810
Ext X Dim	2	4250.82	2125.41	2164.94*	0.1390
Procedure	1	89.51	89.51	91.18*	0.0029
Dim X Proc	2	74.07	37.03	37.73*	0.0024
Ext X Proc	1	63.70	63.70	64.89*	0.0021
Ext X Dim X Proc	2	64.91	32.45	33.06*	0.0021
Sample	1099	5363.44	4.88	4.97*	0.1754

* $p < 0.0001$

In identifying total outliers, the choice of outlier identification procedures was not an issue; although statistically significant, the choice of procedure accounted for less than three-tenths of 1% of the total variability. The degree of extremeness of the outliers accounted for most of the variability in the model; and the dimensionality accounted for a significant amount. There was a significant difference between the two degrees of extremeness and between the two outlier identification procedures. With a minimum significant difference of 0.0495, the two-dimension group was significantly different from the one- and three-dimension groups; the one- and three-dimension groups were not significantly different. Values obtained for Tukey's HSD can be found in Table 7.

Table 7
Tukey's HSD for Total Outliers

Dimensionality	Mean*		
2	A	9.61045	Minimum significant
1	B	8.55727	difference = 0.0495
3	B	8.55432	

* Means with the same letter are not significantly different

An inspection of several of the data sets showed that both outlier identification procedures selected all the induced outliers in the six standard deviation group and in the two and three dimension groups of three standard deviations. The two procedures selected the same observations as outliers in the one dimension three standard deviation group. The one dimension three standard deviation group is the only place in which the procedures did not find all the induced outliers. Table 8 lists the means and standard deviations for the total number of outliers identified by the procedures in the various combinations of dimensionality and extremeness.

Table 8
Means and Standard Deviations for Total Outliers

Procedure	Dimension	Extremeness	Mean*	Standard Deviation
Mahalanobis	1	3σ	8.22	2.28
Andrews-Pregibon	1	3σ	8.93	2.10
Mahalanobis	2	3σ	10.94	1.19
Andrews-Pregibon	2	3σ	11.06	1.24
Mahalanobis	3	3σ	9.05	0.90
Andrews-Pregibon	3	3σ	9.13	0.95
Mahalanobis	1	6σ	8.52	0.70
Andrews-Pregibon	1	6σ	8.56	0.73
Mahalanobis	2	6σ	8.20	0.40
Andrews-Pregibon	2	6σ	8.24	0.43
Mahalanobis	3	6σ	8.02	0.14
Andrews-Pregibon	3	6σ	8.02	0.14

* All means are based on *n*'s of 1,100

BEST COPY AVAILABLE

Conclusions and Discussion

Conclusions

With the two procedures in this study, the Mahalanobis distance and the Andrews-Pregibon statistic, there was a significant difference at the 0.0001 level in detecting both total outliers and false outliers; however, statistical significance was anticipated due to the large number of samples. The two procedures used in this study produced similar results. The squared distance from the centroid (the Mahalanobis distance) and the volume of the confidence ellipsoid with the observation deleted (the Andrews-Pregibon ratio) are both measures of the distance of the observation from the center of the multivariate distribution. In terms of practical significance, the outlier identification procedure accounted for less than 1% of the variability in the model; therefore, other than ease of calculation, there is no reason to choose one procedure over the other procedure. Some statistical packages, such as SPSSX, calculate the Mahalanobis distance, while no commercial package calculates the Andrews-Pregibon statistic. However, using SAS/IML (SAS Institute, 1990b), both procedures can be calculated quickly and efficiently.

The extremeness of the outliers, three or six standard deviations, accounted for 47.3% of the variability using false outliers as the dependent variable and 18.1% of the variability using total outliers as the dependent variable. Both degrees of extremeness were found to be significant at the 0.0001 level; there was a significant difference between the two degrees of extremeness in the number of outliers detected. With either false outliers or total outliers as the dependent variable, the three standard deviation degree of extremeness showed an honestly significant difference from the six standard deviation degree of extremeness. The mean number of outliers identified in the three standard deviation group was higher than that in the six standard deviation group for both false and total outliers.

The dimensionality, outliers occurring in one, two, or three dimensions, accounted for 18.6% of the variability using false outliers as the dependent variable and 10.7% of the variability using total outliers as the dependent variable. All three of the dimensions were found to be significantly different at the 0.0001 level. There was a significant difference among the one-, two-, and three-dimension groups in the number of false outliers detected. There was an honestly significant difference between the two-dimension group and the one- and the three-dimension groups in the number of total outliers detected.

Discussion

It is important to realize that an "outlier" as identified by these two procedures is simply an observation that does not fit the distribution of the other scores. It must be verified that the observations which are identified are true outliers based on an error of measurement, recording, coding, or a deviation from the design.

The choice of an outlier detection procedure does not seem to be a major consideration for the researcher in a situation similar to what was modeled here. It is important for the researcher to be familiar with the data being analyzed and with the procedure being used. He or she must be able to take the observations that are identified as possible outliers by the selected procedure and to examine those observations in order to determine their validity. The decision of whether to accommodate or to reject an observation identified as an outlier must be a careful one. Both procedures in this study identified valid data points as outliers. The presence of these false outliers underscores the necessity of examining the outliers on an individual basis in order to verify that the observation is, in fact, an outlier. Deleting or weighting an observation strictly on the basis of an outlier identification procedure could lead to a loss of valid data. Accepting the results of a procedure without further investigation could lead to invalid results and inferences.

Both procedures identified data points three standard deviations from the mean as outliers. Although the probability of an observation three standard deviations from the mean occurring in normally distributed data is small, the observation could still be from the same distribution; in fact, not having any observations three standard deviations from the mean could signal a problem with the data. Three and six standard deviations were chosen as the degrees of extremeness for this study. In retrospect, three standard deviations was probably not the best choice.

Recommendations for Further Research

Both of the procedures studied were based on distance, distance from the centroid and volume of the confidence ellipsoid. Procedures based on different criteria should be studied as they might produce different results. Many procedures are suggested in the literature; several graphical procedures are available. The decision of which procedure to use should be based on the researcher's knowledge of the procedure, the computer resources available, and the type of analysis to be done. Several procedures are tailored to specific types of analyses, such as regression; therefore, if the researcher has a certain analysis in mind, he or she should investigate the procedures available in that area.

Studies should be done using data involving more dimensions. Procedures are fairly straightforward when there are only two dimensions, but the results become more obscured as the number of dimensions increases.

Studies should be done using different degrees of extremeness. A point six standard deviations from the mean is an outlier, but a point only three standard deviations from the mean is probably a valid data point.

An operational definition of an outlier is needed to facilitate research in this area. In the literature there is no consensus definition of an outlier, much less an operational definition. If the definition must be specified for each study done, the number of available procedures will continue to mount and comparison of results will be difficult. If a common operational definition could be developed, researchers could concentrate on the development of more adequate procedures, possibly through the refinement of existing procedures.

References

Andrews, D. F., & Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society B, 1*, 85-93.

Barcikowski, R. S., & Robey, R. R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. *The American Statistician, 38*, 148-150.

Barnett, V. (1978). The study of outliers: Purpose and model. *Applied Statistics, 3*, 242-250.

Barnett, V., & Lewis, L. (1978). *Outliers in statistical data*. Chichester: John Wiley & Sons.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis in linear regression*. New York: John Wiley & Sons.

Comrey, A. L. (1985). A method for removing outliers to improve factor analytic results. *Multivariate Behavioral Research, 20*, 273-281.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics, 21*, 488-506.

Douzenis, C., & Rakow, E. A. (1987). *Outliers: A potential data problem*. Memphis, TN: Memphis State University. (ERIC Document Reproduction Service No. ED 291 798)

Elashoff, J. D., & Elashoff, R. M. (1970). *A model for quadratic outliers in linear regression*. Stanford University, CA: Stanford Center for Research and Development in Teaching. (ERIC Document Reproduction Service No. ED 047 020)

Gentleman, J. F., & Wilk, M. B. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics, 31*, 387-410.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: John Wiley & Sons.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics, 28*, 81-124.

Gray, J. B. (1989). On the use of regression diagnostics. *The Statistician, 38*, 97-105.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*, 1-21.

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spurosity - A Bayesian approach. *Technometrics, 15*, 723-738.

Hawkins, D. M. (1980). *Identification of outliers*. New York: Chapman and Hall.

Huber, P. J. (1977). *Robust statistical procedures*. Philadelphia: Society for Industrial and Applied Mathematics.

International Business Machines. (1988a). *VM/XA system product interpreter: Reference manual*, First Edition, March 1988: Author.

International Business Machines. (1988b). *VM/XA system product interpreter: User's guide*, First Edition, March 1988: Author.

Jarrell, M. G. (1991, November). *Generating an empirical probability distribution for the Andrews-Pregibon statistic*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, KY.

Krzanowski, W. J. (1988). *Principles of multivariate analysis*. Oxford: Clarendon Press.

Looney, S. W., & Stanley, W. B. (1989). Exploratory repeated measures analysis for two or more groups. *The American Statistician, 43*(4), 220-225.

Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. *Educational and Psychological Measurement, 35*, 707-710.

Pascale, P. J. & Lovas, C. M. (1976). A computer program for detection of statistical outliers. *Educational and Psychological Measurement, 36*, 209-211.

Portnoy, S. (1988). *Regression quantile diagnostics for multiple outliers*. Unpublished manuscript. University of Illinois.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D squared and Comrey Dk. *Multivariate Behavioral Research, 23*, 189-202.

Robertson, C. (1987). The detection of outliers in free recall lists: An exploratory analysis. *British Journal of Mathematical and Statistical Psychology, 40*(2), 140-156.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: John Wiley & Sons.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Rejoinder to Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 648-651.

SAS Institute Inc. (1990a). *SAS procedures guide*, Version 6, Third Edition. Cary, NC: Author.

SAS Institute Inc. (1990b). *SAS/IML software: Usage and reference*, Version 6, First Edition. Cary, NC: Author.

SAS Institute Inc. (1990c). *SAS/STAT user's guide: Volume 2, GLM-VARCOMP*, Version 6, Third Edition. Cary, NC: Author.

Schwager, S. J., & Margolin, B. H. (1982). Detection of multivariate outliers. *The Annals of Statistics, 10*, 943-954.

Seber, G. A. F. (1984). *Multivariate observations.* New York: John Wiley & Sons.

Shapiro, S. S., Wilk, M. B., & Chen, M. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association, 63*, 1343-1372.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*, 334-344.

Tabachnick, B. G. & Fidell, L. S. (1983). *Using multivariate statistics.* New York: Harper & Row, Publishers.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics, 1*, 285-312.

Wood, F. S. (1983). *Measurements of observations far-out in influence and/or factor space.* Paper presented at the Econometrics and Statistics Colloquium at the Chicago Graduate School of Business, Chicago, IL.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 59-62

Gender Differences in Achievement Scores on the Metropolitan Achievement Test-6 and the Stanford Achievement Test-8

John R. Slate, Craig H. Jones, Rose Turnbough, and Lynn Bauschlicher

The presence of gender differences was investigated in secondary students' scores on the Metropolitan Achievement Test (MAT), and elementary and secondary students' scores on the Stanford Achievement Test (SAT). Subjects were 844 Caucasian students enrolled in two schools located in the Mississippi Delta. On the MAT, females exhibited significantly higher scores than did males on the Composite Battery, Total Reading, and Total Language. On the SAT, females obtained significantly higher scores than did males on the Basic Battery, Total Reading, and Total Mathematics. Males did not obtain significantly higher achievement scores than did females on any scales. These results are consistent with other recent research in which females out-performed males in mathematics and support the contention that gender differences favoring males are diminishing.

Maccoby and Jacklin (1974), in their seminal text on psychological differences between the genders, reported that females showed a slight advantage over males in verbal ability, whereas males showed a slight advantage over females in mathematics and spatial abilities. These differences, however, were not found prior to adolescence. Recently, however, researchers have found that the gender differences in both verbal and mathematical ability have virtually disappeared (Hyde & Lynn, 1988; Jacklin, 1989; Marsh, 1989). Only differences in spatial abilities still occur reliably between males and females, and these differences now appear to develop before adolescence (Johnson & Meade, 1987).

The original gender differences reported by Maccoby and Jacklin (1974) raised the issue of whether these ability differences translated into gender differences in academic achievement. Although numerous studies have been published, the results have been highly inconsistent. For example, in a review of 15 studies, Steinkamp and Maehr (1983) concluded that males show slightly better science achievement than do females. Similarly, in The Gender Gap (1989) study, females were reported to out-perform males slightly in reading and writing, whereas males out-performed females slightly in mathematics and science. Other investigators (e.g., Randhawa & Hunt, 1987; Shaw & Doan, 1987), however, have found no differences in science achievement between males and females. Friedman (1989) conducted a meta-analysis of 98 investigations and found that the gender difference in science favoring males declined as a function of the year in which the study was conducted. More recently, Randhawa (1991) reviewed studies conducted from 1978 to 1989 and found a gender difference favoring females in language and, surprisingly, some evidence for a slight advantage in mathematics for females as well.

One problem with previous research on gender differences in academic achievement has been the use of nonstandardized measures (e.g., high school examinations) to measure academic achievement. Notable exceptions are studies by Scott (1987), Shaw and Doan (1990), and Hayes and Slate (1993). Even these studies, however, have produced inconsistent results. Shaw and Doan (1990) found no gender differences in science on the Stanford Achievement Test, but Scott (1987) found that females exhibited higher achievement on all subtests of the California Achievement Test. Hayes and Slate (1993) found that high school females obtained higher scores than did high school males on the Metropolitan Achievement Test.

Given both the changing findings with regard to gender differences in intellectual abilities and the inconsistent results with regard to gender differences in academic achievement, additional research using standardized measures of academic achievement is needed. In this article, the results of two studies are reported. In the first study, gender differences in the academic achievement of secondary school students on the sixth edition of the Metropolitan Achievement Test (MAT-6) were examined. Study 1, therefore, was a replication of research by Hayes and Slate (1993) conducted to determine if gender

---

John R. Slate and Craig H. Jones are Professors in the Department of Counselor Education and Psychology at Arkansas State University. Rose Turnbough is a counselor in the Mammoth Springs Schools in Mammoth Springs, Arkansas. Lynn Bauschlicher is a counselor in the Corning Public Schools in Corning, Arkansas. Please address correspondence regarding the paper to John R. Slate, Ph.D., Department of Counselor Education and Psychology, Arkansas State University, PO Box 940, State University, AR 72467-0940.

differences are present in the Composite Battery, Total Reading, Total Math, Total Language, Science, and Social Studies scores on the MAT-6 for students in grade 7 through grade 12. The second study was a replication of research by Shaw and Doan (1990) conducted to determine (a) if gender differences are present in the Composite Battery, Total Reading, Total Math, Total Language, Science, and Social Studies scores on the SAT-8 for students in grade 1 through grade 6, and (b) if gender differences are present in the Composite Battery, Total Reading, Total Math, Total Language, Science, and Social Studies scores on the SAT-8 for students in grade 7 through grade 11.

Study 1

Method

Data were collected on 481 students (230 females, 251 males) in grades 7 through 12 enrolled in a rural school district in northeast Arkansas. These students had completed the MAT-6 in the spring of 1991. The sample was exclusively white. This school was located in the Mississippi Delta, indicating that the majority of students came from lower class socioeconomic backgrounds. In fact, 36% (35% of females and 37% of males) of our sample qualified for either free lunch/breakfast (29%) or reduced price lunch/breakfast (7%) in the National School Lunch/Breakfast Program.

Data obtained from students' permanent records included their Composite Battery, Total Reading, Total Mathematics, Total Language, Science, and Social Studies percentile scores on the MAT-6. Similar to the Hayes and Slate (1993) study, these percentile scores were converted to a standard score format for statistical analysis for two reasons. First, the means and standard deviations for each scale differ by grade level and content area. Conversion to standard scores produced a uniform mean of 100 and standard deviation of 15 for all grade levels and content areas. Second, the conversion to standard scores permitted statistical analyses that could not be conducted on the ordinal level data provided by percentile scores. Thus, the standard scores and not the percentile scores were subjected to the statistical analyses conducted in this study.

Results

The means and standard deviations for each gender on the Composite Battery and the five subscales are displayed in Table 1. Females obtained higher means than did males on five of these six measures. Only on the Science subscale was the average male score higher than the average female score. An analysis of variance of the Composite Battery scores, $F(1, 477) = 3.85$, $p < .05$, revealed that females ($M = 103.9$) demonstrated significantly higher overall academic achievement than did males ($M = 101.5$). A multivariate analysis of variance of the scores for Total Reading, Total Mathematics, Total Language, Science, and Social Science was also statistically significant, $F(5, 467) = 12.04$, $p < .001$, indicating the presence of gender differences in these subscales. Univariate analyses of variance revealed that females ($M = 102.5$) scored significantly higher than did males ($M = 99.8$) on Total Reading, $F(1, 471) = 5.07$, $p < .05$. Females ($M = 107.7$) also significantly outscored males ($M = 101.6$) on Total Language, $F(1, 471) = 26.74$, $p < .001$). Statistically significant differences were not found between females and males on Total Math, $F(1, 471) = 2.31$, Science, $F(1, 471) = 0.20$, or Social Studies, $F(1, 471) = 0.35$.

Table 1
Means and Standard Deviations of Males and Females
in Grades 7 Through 12 on the MAT-6 Scales

	Females ($n = 230$)		Males ($n = 251$)	
Scale	Mean	SD	Mean	SD
Composite Battery	103.9	12.7	101.5	14.4
Total Reading	102.5	12.9	99.8	14.9
Total Mathematics	104.6	12.6	103.0	14.1
Total Language	107.7	12.4	101.6	13.4
Science	102.1	11.8	102.6	13.9
Social Science	101.9	13.4	101.3	15.11

Study 2

Method

Data were collected on 363 students in grade 1 through grade 11 enrolled in another rural school district in northeast Arkansas. These students completed the SAT-8 in the spring of 1992. The SAT-8 was not administered to students in the 12th grade at this school. There were 193 elementary students (91 females; 102 males) and 170 secondary students (88 females and 82 males). This sample was also exclusively white. Because this school was also located in the Mississippi Delta, the majority of students again came from lower socioeconomic class backgrounds with 65% (66% of females and 64% of males) qualifying for either free or reduced prices in the National School Lunch/Breakfast Program.

Data obtained from students' permanent records included the following SAT-8 scores: Composite Battery, Total Reading, Total Mathematics, Total Language, Science, and Social Studies. Because SAT-8 means and standard deviations also differ by grade level and content area, percentile scores were again converted to standard scores with a mean of 100 and standard deviation of 15. Gender differences were examined separately for the elementary grades (i.e., 1-6) and the secondary grades (i.e., 7-11).

Results

Elementary grades. The means and standard deviations for each gender on the Composite Battery and the five subscales are displayed in Table 2. Analysis of variance indicated that males and females did not differ significantly on the Composite Battery, $F(1, 191) = 0.09$, indicating that males and females did not differ in their overall academic achievement. A multivariate analysis of variance of the five subscale scores was also nonsignificant, $F(5, 135) = 1.35$, indicating that male and female achievement did not differ as a function of the specific academic area as well.

Table 2

Means and Standard Deviations of Elementary School Males and Females in Grades 1 Through 6 on the SAT-8 Scales

	Females ($n = 91$)		Males ($n = 102$)	
Scale	Mean	SD	Mean	SD
Composite Battery	104.1	11.4	103.5	13.3
Total Reading	102.5	11.7	102.6	13.3
Total Mathematics	108.1	11.3	107.8	13.7
Total Language	100.9	10.4	98.9	11.9
Science	104.1	12.8	105.9	13.3
Social Science	102.4	11.3	102.4	13.1

Secondary grades. The means and standard deviations for each gender on the Composite Battery and the five subscales are displayed in Table 3. Analysis of variance indicated that males and females differed significantly on the Composite Battery, $F(1, 168) = 4.10$, $p < .05$, with females ($M = 102.4$) demonstrating higher overall achievement than did males ($M = 98.0$). A multivariate analysis of variance of the five subscale scores was also significant, $F(5, 77) = 3.22$, $p < .01$, indicating that male and female achievement also differed as a function of specific academic areas. Univariate analyses of variance revealed that females ($M = 101.8$) demonstrated higher achievement than did males ($M = 97.9$) on Total Reading, $F(1, 169) = 3.64$, $p < .05$. Surprisingly, females ($M = 100.9$) also demonstrated significantly higher achievement than did males ($M = 96.2$) on Total Mathematics, $F(1, 168) = 3.66$, $p < .05$. Significant differences were not found for Total Language, $F(1, 81) = 1.47$, Science, $F(1, 169) = 0.32$, or Social Science, $F(1, 69) = 0.01$.

Table 3

Means and Standard Deviations of Secondary School Males and Females in Grades 7 Through 11 on the SAT-8 Scales

	Females ($n = 88$)		Males ($n = 82$)	
Scale	Mean	SD	Mean	SD
Composite Battery	102.4	13.8	98.0	13.5
Total Reading	101.8	13.3	97.9	13.4
Total Mathematics	100.9	14.1	96.2	13.8
Total Language	98.9	12.6	95.8	11.2
Science	101.9	12.8	100.8	12.5
Social Science	100.8	12.2	100.9	13.3

Discussion

Although no gender differences were found in elementary school children's academic achievement, secondary school females demonstrated higher overall academic achievement than did males on both the MAT-6 Composite Battery and the SAT-8 Composite Battery. Females also out-performed males in Total Reading on both the MAT-6 and the SAT-8, in Total Language on the MAT-6 only, and, in Total Mathematics on the SAT-8 only. Males did not out-perform females on any of the achievement measures to a statistically significant degree.

These findings agree with Friedman's (1989) contention that gender differences favoring males are diminishing. In addition, higher scores by females in mathematics on the SAT-8 are consistent with other recent research in which females out-performed males in mathematics (Hayes & Slate, 1993; Randhawa, 1991). Females, on the other hand, continued to out-perform males in at least some areas of language achievement.

For example, females exhibited better reading skills than did males in both studies. These superior reading skills may at least partially explain females having higher levels of overall academic achievement than do males.

These results are based upon students from only two schools in a geographically restricted area (i.e., the Mississippi Delta). Moreover, all subjects were Caucasian with low social class backgrounds. In fact, 36% and 65% of our samples qualified for free or reduced price lunches, thus meeting the criteria for poverty. Therefore, generalizations must be made only with considerable uncertainty. As noted above, however, the findings are also reasonably consistent with the results of several other recent studies. Thus, the possible implications of these findings do need to be discussed.

Recently, social concerns have been raised that girls are not performing well relative to boys in mathematics and science (High School, 1992). The results of this study support the growing consensus among researchers that such differences probably do not exist, or, that any differences which do exist are small (Friedman, 1989). Even in our study, the largest gender difference was only 6 points in Total Language on the MAT-6. The public concern noted above appears to be based on the original Maccoby and Jacklin (1974) studies rather than on more recent research. Thus, better dissemination of recent findings to the public, especially policy makers, is needed.

From a research standpoint, many researchers now believe that gender differences between males and females have largely dissipated. In the present study, secondary school females not only out-performed males on a number of measures of verbal achievement, but also exceeded males on overall achievement (both SAT-8 and MAT-6) and in mathematics (SAT-8). Randhawa (1991) also reported a slight advantage in mathematics for females. This raises two questions for researchers. First, to what extent do the small but statistically significant gender differences being found on standardized achievement tests translate into meaningful academic differences? Second, have previous differences in academic achievement between males and females simply closed, or are females slowly coming to out-perform males even in those areas in which males traditionally have been expected to out-perform females? Additional studies employing both standardized test scores and more "authentic" measures of achievement will be needed to answer these questions.

References

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematics tasks. *Review of Educational Research, 59*, 185-213.

Hayes, L., & Slate, J. (1993). Differences in MAT6 test scores by gender. *Louisiana Educational Research Journal, 18*, 161-166.

High school boys out-perform girls on science tests. (1992, March 27). *Arkansas Democrat-Gazette*, p. 4.

Hyde, J., & Lynn, M. (1988). Are there sex differences in verbal abilities?: A meta-analysis. *Psychological Bulletin, 104*, 53-69.

Jacklin, C. (1989). Female and male: Issues of gender. *American Psychologist, 44*, 127-133.

Johnson, E., & Meade, A. (1987). Developmental patterns of spatial ability: An early sex difference. *Child Development, 58*, 725-740.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Marsh, H. (1989). Sex differences in the development of verbal and mathematical constructs: The High School and Beyond Study. *American Educational Research Journal, 26*, 191-225.

Randhawa, B. (1991). Gender differences in mathematics: A closer look at mathematics. *Alberta Journal of Educational Research, 37*, 241-257.

Randhawa, B., & Hunt, D. (1987). Sex and rural-urban differences in standardized achievement scores and mathematics subskills. *Canadian Journal of Education, 12*, 137-151.

Scott, R. (1987). Gender and race achievement profiles of black and white third-grade students. *The Journal of Psychology, 121*, 629-634.

Shaw, E., & Doan, R. (1990, April). *An investigation of the differences in attitude and achievement between male and female second and fifth grade science students*. Paper presented at the 63rd Annual Meeting of the National Association for Research in Science Teaching, Atlanta, GA.

Steinkamp, M., & Maehr, M. (1983). Affect, ability, and science achievement: A quantitative synthesis of correlational research. *Review of Educational Research, 53*, 369-396.

The Gender Gap in Education: How Early and How Large? (1989). *ETS Policy Notes, 2*(1), 1-9.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 1, 63-76

The Global Coherence Context in Educational Practice: A Comparison of Piecemeal and Whole-Theme Approaches to Learning and Teaching

Asghar Iran-Nejad

Recent biofunctional research suggests that global coherence is a natural aspect of normal brain functioning (the global coherence assumption). This paper argues that another assumption in educational research and practice, the assumption of simplification by isolation, runs counter to the implications of the global coherence assumption and is the source of many of the problems in today's education. Whereas the global coherence assumption implies a whole-theme approach to learning and teaching, simplification by isolation pushes educational research and practice toward a piecemeal approach. This paper compares and contrasts piecemeal and whole-theme approaches and concludes that a whole-theme approach can potentially enable us to rethink our existing ways of going about learning, teaching, and organizing learning environments in a radically different fashion.

One of the greatest challenges of education is to teach today's learners to function in the real world of tomorrow. To think of preparing students now for more than two decades into the future in the rapidly-changing context of the modern world is overwhelming, especially when education is receiving failing grades these days for preparing students for the real world of today (Bigler & Lockard, 1992; Meyers, 1986). Even the positive effects of formal education have not always firmly withstood the power of closer scrutiny. For instance, Voss, Blais, Means, Greene, and Ahwesh (1989) found that performance differences between naive and novice learners in economics (those without and with formal training in this area, respectively) disappeared when measures matched less closely the formal education concepts and more closely the economic issues of every-day life. This paper argues that the failure of today's education (e.g., to produce authentic learning) may be largely a result of the fact that educational theory and practice are caught in the unrelenting grip of what Bartlett (1932) called the assumption of simplification by isolation.

Recent research on conceptions of learning (Marton, 1988) suggests that they exert a profound influence on the way we approach schooling (Iran-Nejad, 1990). The assumption of simplification by isolation has been identified as one of the most counterproductive roots of these tacit conceptions (Iran-Nejad, McKeachie, & Berliner, 1990).

As researchers, teachers, or learners, we tend to think that complex tasks become more manageable (i.e., easier) once broken down into their so-called basic components. The result is an everlasting shift in educational practice away from what has authentic (real-world) relevance, because the real world is highly complex by nature (Schon, 1987), and toward isolated skills, facts, concepts, or principles--regarded as prerequisite knowledge for complex real-world problem solving. Thus, the gap between what Schon (1987) called the stone-solid hill of professional knowledge and the slimy-soft swamp of real-world problem solving continues to widen. Accumulation of basic-level knowledge becomes the business of today's education and complex real-world problem solving is left forever to be a topic for the future.

Consider how the assumption of simplification by isolation might work in schools. Under its influence, a teacher may believe, and practice accordingly, that more elementary and lower ability classes must focus on the teaching of isolated concepts, facts, and principles (Shuell, 1990). The teaching of the so-called higher-order thinking, on the other hand, which would require as prerequisite considerable accumulation of low-level knowledge, must be reserved for advanced and high ability learners who have already amassed the basic stuff. One unfortunate consequence of this hierarchical assumption is that for several decades, low-level-knowledge teaching has cast its shadow over the field of education with the total exclusion of higher-order thinking (Bloom, 1984). This realization has brought about the recent widespread appeal by researchers and practitioners for an active and conscious focus on higher-order teaching (Newmann, 1990; Peterson, 1988; Prawat, 1989) with a positive influence in guiding the direction of thinking in education toward authentic learning and problem solving.

Asghar (Ali) Iran-Nejad is an Associate Professor of Educational Psychology in the College of Education at The University of Alabama. Correspondence and requests for reprints should be addressed to Dr. Asghar Iran-Nejad, Educational Psychology, The University of Alabama, P. O. Box 870231, Tuscaloosa, AL 35487-0231.

This is no doubt an important development. Equally important, it seems, are (a) a commitment to eradicating the counterproductive assumptions underlying existing educational practices and (b) replacing these assumptions with productive alternatives. Widespread change in the traditional culture of schooling is unlikely so long as the tacit root-level assumptions that drive educational practice on a day-to-day basis remain intact:

> Despite the emerging consensus on the importance of teaching for higher-order thinking, research . . . generally finds that classroom instruction in high schools is focused on basic skills (Goodlad, 1984; Powell, Farrar, & Cohen, 1985). To the extent that teaching for higher-order thinking is manifest, evidence suggests that it occurs far more often in high-track than low-track classes (Metz, 1978; Oakes, 1985; Page, 1990). Thus, at high school level in the United States, a sharp contrast exists between current visions of educational excellence and currently institutionalized patterns of educational practice. (Raudenbush, Rowan, Cheong, 1993, p. 524)

There appears to be a vast imbalance on the degree of emphasis in favor of basic skills at the expense of higher-order problem solving and critical thinking (Hannaway, 1992). Raudenbush et al. asked teachers to report the amount of emphasis they placed on higher-order thinking in their classes. The study involved 16 high schools in California and Michigan using a sample of 303 teachers and 1,205 classes in math, science, social studies, and English. The results showed a powerful influence of conceptions of learning, particularly true of math and science but also significant for social studies and English. Teachers tended to observe, to use the terminology of the present discussion, a simplification-by-isolation focus in more elementary and lower-track classes and reserve higher-order thinking objectives for high-track students in advanced courses. If, as mentioned already and suggested by the Raudenbush et al. data, it is indeed the deep-seated assumptions behind conceptions of learning that hold down firmly the roots of the currently institutionalized culture of educational practice, the conscious willingness or decision to focus on higher-order thinking is unlikely to solve educational problems single-handedly. Raudenbush et al. looked at the influence of teacher training and background experience. Their results "provided no evidence that simply having obtained a master's degree or having extensive teaching experience predisposed a teacher to pursue higher-order objectives. Rather, the match between the teacher's preparation and the subject matter of a particular class appeared to be linked to higher-order emphasis" (p. 548). The link with the subject matter knowledge is not surprising because the more fluent teachers are in the subject matter they teach, the more likely they are to focus spontaneously--perhaps even without knowing--on critical thinking in that domain. Neither surprising is the absence of any effects of the educational background or experience of the teacher, given the notion that the roots of simplification-by-isolation are "deeply institutionalized in conceptions of teaching and learning that are essentially invariant across teachers and organizational environments" (p. 528), including those that govern teacher training and classroom teaching practices.

The implications of this line of reasoning for school restructuring and educational reform efforts are obvious. Innovative reform structures "in and of themselves are not necessarily associated with higher levels of classroom thoughtfulness" (Ladwig & King, 1992, p. 710). So long as reform efforts remain focused at the shallow level of active and conscious decision making and leave the entrenched root-level conceptions of teaching and learning untouched, they are unlikely to change substantially the direction of educational practice. What is needed is a direct, systematic, and root-level attack by means of rigorous teacher training programs on the problems of identifying, exposing, and replacing counterproductive assumptions. The root-level strategy implies that reform, restructuring, or rethinking the process of education cannot be attained by telling, requiring, making accountable, or tightening standards. All of these are shallow-level measures. Root-level measures are likely to require the development of long-term teacher training programs consisting of both a preteaching training period aimed at uprooting and replacing counterproductive conceptions of learning and an in-school training aimed at helping teachers-in-training to implement their preteaching knowledge to build a radically different school culture.

No one knows at the present time how many years of preteaching or on-the-job training it takes to develop such a radically different school culture. I suspect that it is going to require more time than is currently assigned to most traditional teacher education programs. Neither is it possible to plan in detail what teacher training programs ought to cover. What is certain is that if alternatives are found to such assumptions as simplification by isolation and are implemented successfully, the emerging school culture will be radically different, and more promising. Consistent with this notion, this paper makes the following assumptions: (a) the assumption of simplification by isolation is a major, widespread, and seductive cause of existing educational problems; (b) simplification by isolation is not the only way to simplify learning situations; (c) an alternative root-level assumption to simplification by isolation is the opposite notion of simplification by integration; and (d) extensive theoretical and practical effort is required to reevaluate, rethink, and reorganize institutionalized conceptions of learning by uprooting the assumption of simplification by isolation in all its manifestations and replacing it with the radically different but promising assumption of simplification by integration. It is in the spirit of this set of assumptions that the present paper compares and contrasts what is called hereafter the piecemeal approach to teaching and learning, caused primarily by the assumption of simplification by isolation, with a dramatically different whole-theme approach, based on the opposite assumptions of simplification by reorganization and integration.

To get a feel for how the traditional (piecemeal) and the alternative (whole-theme) approaches are different in terms of their expected teacher-training outcomes, consider a thought experiment. Imagine a sample of schools similar to that used by Raudenbush et al. (1993), divided into two subsamples, and randomly assigned to a traditional (piecemeal) or a whole-theme teacher training approach. The traditional sample is then taught in the way teachers are traditionally trained. The whole-theme sample is taught in a training program comparable in content and duration but different in approach. Subjects in the traditional sample are expected to continue to allocate higher-order objectives only to high grade/rank classes as did the subjects in the Raudenbush et al. experiment; but subjects in the whole-theme sample will tend to allocate higher-order objectives independently of grade/rank.

A Whole-Theme Analysis of the Piecemeal and Whole-Theme Approaches

The Domain-Launching Theme (DLT)

Suppose that we are planning a course called Learning and Teaching for School Teachers to be taught to subjects in the above thought experiment. The first step in the traditional school-culture approach is to break the content of the course into its many components and place them in a sequential order for presentation. By contrast to this piecemeal way of schooling, the very first step in preparing for and teaching the course from a whole-theme perspective is to develop a domain-launching theme (DLT) for it and translate this theme into a tangible DLT organizer, one that can be consulted and used again and again by all students throughout the course. A DLT organizer is an external thematic organizer depicting the entire domain of learning and teaching for school teachers. This thematic organizer is used to introduce, as the very first step in teaching, the entire domain of the course and its content to the students. This is the central idea behind the whole-theme approach: to present the entire domain of a course or subject matter all at once right at the outset in the form of a single external representation of the course.

In the piecemeal approach, the entire domain comes towards the end, if ever, after all the prerequisite parts have already been internalized piece by piece. In the whole-theme approach, a total-picture of the entire domain must come first in the form of a DLT organizer. This is the simplification-by-reorganization alternative, because the goal is to change the organization of the learner's existing knowledge base into the new organization anchored by the DLT organizer. Moreover, the remainder of the course must be organized in sequence such that the course content is learned and/or taught directly in the form of problems to be solved in the context of the DLT organizer. This is simplification by integration, because the goal is to integrate systematically the details, so to speak, of the learner's existing knowledge base into the new theme established and anchored by the DLT organizer. Since this paper is not about thematic organizers, I will not dwell on them or on the various ideas that go into their building here. Instead, I will illustrate the notion of a thematic organizer with the DLT organizer that I use in teaching a cognitive educational psychology course to graduate students.

Figure 1 shows this DLT organizer. I believe the same thematic organizer can be used to teach learning and teaching to future teachers. The DLT organizer portrays learning as an evolving process of personal growth during which the individual progresses from being a (naive) newcomer to a domain toward an effective professional in that domain. Examples of domains are particular subject matters such as cognitive educational psychology, reading, or physics. The diagonal cone-arrow from lower left to upper right represents piecemeal learning, which extends linearly from no professional knowledge at all to a large store of professional knowledge. The diagonal cylinder-arrow from upper left to lower right represents whole-theme learning during which a naive learner's intuitive knowledge base (represented by the cylinder) evolves into a professional knowledge base (PKB) through an indefinite, but not very large, number of nonlinear thematic reorganizations. Each global ring in the learner's intuitive knowledge base (IKB) is meant to be

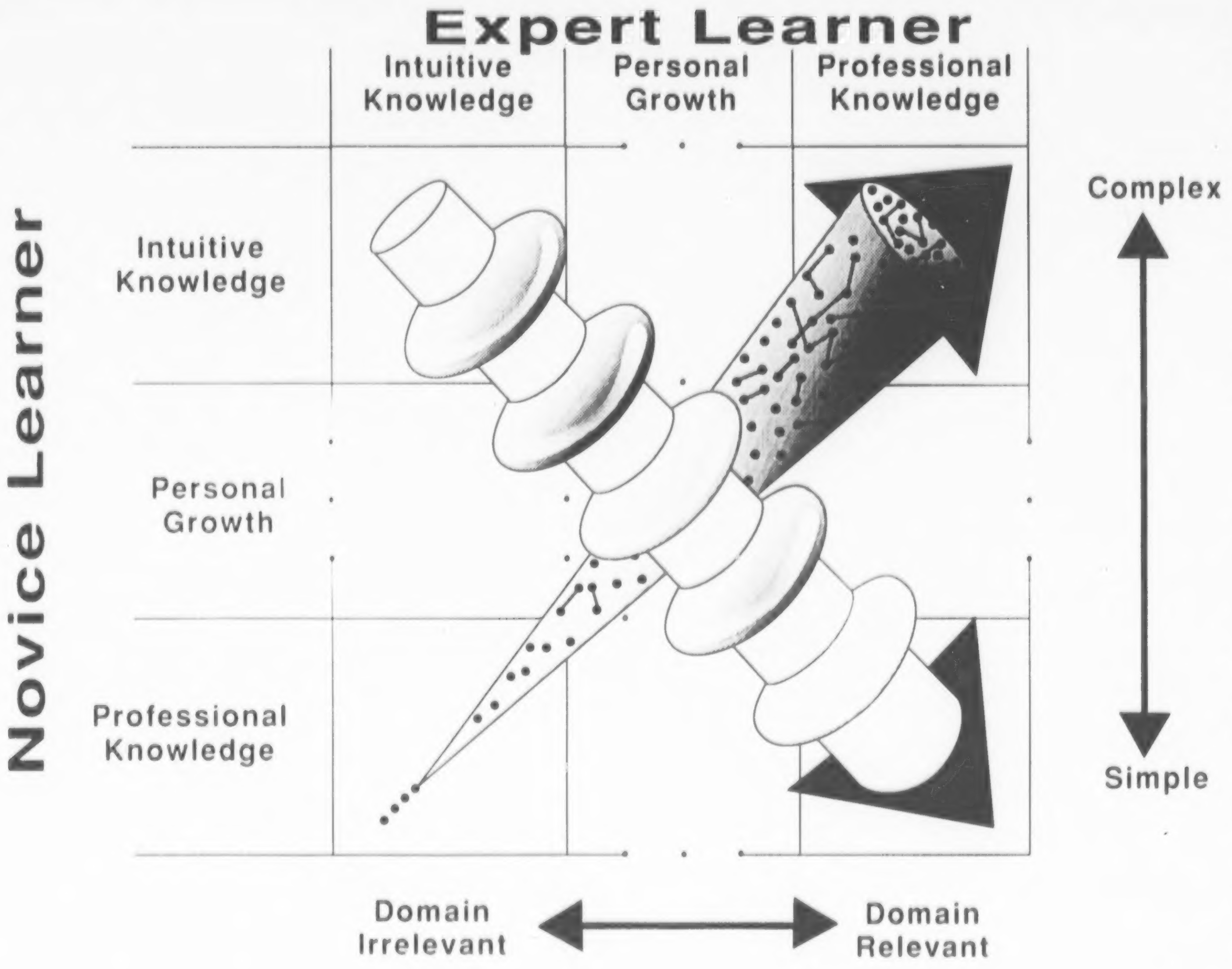

Expert Learner
Intuitive Knowledge
Personal Growth
Professional Knowledge
Novice Learner
Intuitive Knowledge
Personal Growth
Professional Knowledge
Complex
Simple
Domain Irrelevant
Domain Relevant

Figure 1. A thematic organizer for a comparison of the piecemeal and whole-theme approaches to learning and teaching. Each ring on the diagonal from upper left to lower right is meant to be in a different color (red, blue, green, purple, orange, respectively) representing a different theme.

BEST COPY AVAILABLE

in a different color (say, red, blue, green, purple, and orange, respectively) depicting a different theme to serve as a global coherence context for further problem solving toward the establishment of local-level coherence between that theme and the learner's IKB. It might be noted that the PKB is a hypothetical concept in that there is no ultimate PKB. It is more realistic to think of IKB as a dynamic system always in a state of *becoming* a PKB but never reaching the static end state of actually *being* one.

Dictionaries define intuitive knowledge as the knowledge learned directly--without conscious reasoning--through immediate apprehension or understanding. Before school, children demonstrate a remarkable capacity for intuitive learning from experience. They manage to acquire a functional knowledge of the world around them as well as of their mother tongue. It is this knowledge that constitutes their initial IKB. It is this knowledge that the traditional piecemeal school culture ignores by assuming, and starting with, zero knowledge of PKB and proceeding to build up the PKB piece by piece.

Referring back to Figure 1, the piecemeal cone-arrow shows the way students are taught in schools. As the figure suggests, because of their direct and almost exclusive focus on subject-matter knowledge and under the influence of the assumption of simplification by isolation, schools force children to leave their IKB behind day after day as they enter the school and the classroom. The assumption of simplification by isolation is very seductive. In the past three decades, there has been a widespread campaign as well as a great deal of research and evidence supporting the critical role background knowledge plays in learning (Anderson, Spiro, & Montague, 1977; Schoenfeld, 1987; Spiro, Bruce, & Brewer, 1980). More recently, there has been a similar campaign for a focus on higher-order thinking. However, as the research by Raudenbush et al. (1993) suggests, the assumption of simplification by isolation serves as a solid wall resisting firmly the impact of these developments. As a result, as the piecemeal learning arrow in Figure 1 shows, school learning totally bypasses the leaner's IKB, that plays such a fundamental role in her or his daily living (Rovegno, 1993). The piecemeal portion of Figure 1 can go a long way in explaining why the gap between what is learned in schools and what is relevant in the real world of practice continues to widen throughout formal schooling years (Schon, 1987).

Whole-theme learning (cylinder-arrow diagonal) tells a radically different story. First, the learning of the new domain must begin in the real-world-rich context of the learner's IKB. There is a potential advantage here of immense magnitude. In recent years, there has been a major shift in thinking toward the situated nature of learning. This is an important movement because it draws attention to authentic, as opposed to academic, learning. The new movement suggests that situational authenticity must be protected and nourished in educational practice. However, the whole-theme approach implies that situatedness itself can be the very shallow tip of the giant iceberg of the process of authentic learning. For instance, the implicit contrast between situatedness and the general/abstract knowledge makes it easy, though not necessarily so, to go to the other extreme and view situatedness in terms of shallow external scenarios or cases (Hintzman, 1986) at the expense of the vast contribution of the individual variables--variables internal to individual learners themselves. As a result, situatedness often reduces to shallow-level instance learning, or piecemeal accumulation of examples or cases in overly particularized settings. When this happens severe obstacles present themselves in terms of practical applications. It is neither always possible nor desirable to conduct the teaching of complex domains entirely in their particularized real-world settings.

The whole-theme approach implies that situatedness must be defined as situational authenticity first in terms of the real-world-rich IKB within the learner and only then in the form of authentic practice in actual real-world contexts. The whole-theme approach, then, has all of the advantages of the original notion of situatedness without many of its practical limitations. In other words, whole-theme learning promises to capture and formalize the essential spirit of such fundamental approaches as contextualism (Jenkins, 1974), situated cognition (Brown, Collins, Duguid, 1989; Clancey, 1993; Greeno & Moore, 1993), and immersion (Prawat, 1991). The focus of the whole-theme approach, however, is on the role of the naive learner's IKB in learning and its evolution through reorganization toward the expert learner's professional knowledge base (PKB). The basic idea is that the launching of the problem-solving journey toward a new PKB must occur not as a separate domain and not with a few fragmented subskills, concepts, facts, definitions, or procedures but with an IKB-based, real-world-rich, holistic theme.

Problem-Solving in Complex Domains

The whole-theme approach is not a teaching/learning method. Rather, it is an approach to problem solving, especially in highly complex domains such as teaching teachers how to teach. Thus, when taught from the whole-theme approach, Cognitive Educational Psychology or Learning and Teaching for School Teachers are not conventional courses. They are courses in problem solving in their respective domains. Extensive root-level problem solving can be done in the context of whole-theme university courses such as these. Consider asking college of education majors to try to solve the following problem using our DLT organizer and their own IKB: Should the piecemeal and whole-theme approaches depicted in Figure 1 be treated as completely mutually exclusive or is it better to use them together in teaching? This problem was posed to the 12 graduate students enrolled in the Cognitive Educational Psychology course already mentioned. A representative sample of their thoughts on the solution is presented in Table 1.

Many graduate students accept the whole-theme approach soon after the DLT organizer is presented to them. They also agree that it represents a dramatically different perspective from the traditional piecemeal approach. This is perhaps because the DLT organizer serves as a good vehicle for globally reorganizing individual students' IKBs and establishing within them a global coherence context. For those students who have not conceived the traditional and alternative cultures of schooling in the manner portrayed in Figure 1, this is a landmark experience of no small consequence. Establishing a new global-level organization, however, does not mean total IKB reorganization on the part of the students. For instance, many of the cognitive educational psychology students who readily acknowledge that the whole-theme and piecemeal approaches are dramatically different, continue to experience considerable difficulty in accepting the complete mutual exclusion hypothesis, as the data in Table 1 illustrates. This is presumably because many institutionalized root-level assumptions in their IKBs resist total integration into the global coherence theme. For total integration to occur, extensive problem solving is required to reach and reorganize the many long-held beliefs that lie deeply at the very roots of their IKBs.

Authentic Learning in the Classroom

Thus, a relatively large number of problems must be solved by the learner before there can be coherent global as well as local continuity between the DLT and individual student IKBs. Such a process of IKB-DLT integration by means of problem solving in highly complex domains is authentic learning. An important implication of the whole-theme approach, as far as practical applications are concerned, is that much authentic problem solving can occur in the traditional classroom setting in the form of, say, the preteaching training of teachers.

One way authentic learning differs from academic learning, in the traditional sense, is how readily it transfers, or is applied by the learner to new situations. There is some evidence that the whole-theme approach itself, as taught in the Cognitive Educational Psychology course already mentioned, is highly transfer-appropriate. One indication of this became evident in my experience with teaching the course during Spring 1993, which was the first time that I used the present DLT organizer. During the seventh week of the semester, 8 out of the 12 students enrolled in the course included a thematic organizer in their first required essay. Moreover, at least three of the students used the whole-theme approach in their own teaching in the same semester in areas as diverse as undergraduate educational psychology (Cochran, 1993), supervision of student teachers (Volkman & Iran-Nejad, 1993), and tests and measurement (Zheng & Iran-Nejad, 1993). In all three cases, the decision to apply the whole-theme approach was made soon after the introduction of the DLT organizer. Cochran (1993) reports on his teaching experience with the whole-theme approach in the following way: "Acceptance of the whole-theme approach to instruction feels liberating and is not simply the acceptance of another paradigm with rigid boundaries set by others for the teacher to follow. Inherent to the approach is the freedom of creativity. . . . As the teacher or learner, I am not conforming to someone else's template" (p. 10).

These are examples of thematic transfer. Fluent on-the-job application of the whole-theme approach is expected to require widespread integration of the learner's IKB to the DLT organizer. This is by no means a small challenge and is expected to involve extensive preteaching and on-the-job training and experience in problem solving. The present author has had several years of background and experience with the whole-theme approach, much of which has been a rocky road of challenge and gratification which has only recently turned into the kind of fluency one experiences in advanced stages of learning a foreign language.

The most immediate test of transfer-appropriateness of a model is whether or not it can be practiced successfully by its designer. In other words, how readily a model of complex problem solving can be used in new real-life situations by its designer, who presumably knows everything there is about the model, is the minimum requirement for its transfer-

Table 1

Cognitive Educational Psychology Students' Solutions to the Problem: Should Piecemeal and Whole-theme Approaches be Treated as Completely Mutually Exclusive or Is It Better to Use Them Together in Teaching?

Joe Green: In my assumption of *simplification through experimentation*, one would not use either of the two approaches exclusively. There is a certain amount of piece that must come together during the process of integration. The executive committee would in fact take parts from piecemeal approach and parts from the whole-theme approach to form a concrete theory. To start with the whole is a good idea. However, once we have the whole theme or the complex situation, break themes down and take a step by step approach to simplify the situation.

Sue Smith: I think the two approaches are entirely different but I do believe there are times when they might blend. There may be times when it will become necessary to isolate a particular idea or fact for the purpose of emphasis even though you may be using a whole-theme approach. If you use a piecemeal approach, it may still include multi-sensory experiences though to not as great an extent. The piecemeal approach is not as meaningful because it does not present the whole picture but it may, at times, use several sources.

C. S. Lewis: I like the idea of a whole-theme approach. However, I am still having problems completely separating the two. At some point, every learner must give attention to the minute detail and learn what it is. Before whole-theme approach to most any subject can be effective, some isolation of parts must occur. Even in your illustration of learning one's native language, the parents of the child attend to some isolated parts of teaching their child what some things are. E.g., "Ball, this is a ball. Can you say ball?" However, I also see at some point that the multisource/multimodal kicks into overdrive of the language explosion period of the older toddler. Now to specifically address your question, I have difficulty, completely separating the two.

Ariel: I think a whole-theme approach is a better match with *how* people learn than is the piecemeal approach. I think it is possible to present information in a piecemeal fashion *after* the whole-theme is introduced. However, using the piecemeal approach alone leads to a number of problems with learning: lack of meaning or relevance, boredom and distractions, to name but a few. All of these problems may be lessened or alleviated when the theme is presented first, perhaps because as humans we seek meaning in our life experiences naturally and learning in this approach is a natural extension of how our system organizes information. If we have no theme and only get bits of information, we try to make a theme, or build a theme around the bits of information. So the two approaches are not mutually exclusive necessarily--we may use bits and pieces to fill in the big picture (whole-theme) where we have gaps in our intuitive knowledge. Yet without the whole-theme, we will create a theme (even if it is "learn so I can pass the test").

Tony Cole: As I understand piecemeal, bits of information are learned in somewhat of a sequential way until the learner understands the domain he/she is learning. As I understand the whole-theme approach the learner uses his/her intuitive knowledge in a meaningful way and the concepts are changed as more knowledge is gained. It may appear that there is gathering of piecemeal information in the whole-theme approach as more information is gained through multisource means. However, it is not the same as gathering pieces and finally arriving at learning the body of knowledge. In the whole-theme approach, the learner uses his/her intuitive knowledge and applies the information to revise concepts or to make accommodations in how he/she views and understands the body of knowledge. The piecemeal approach is not really used in the whole-theme approach.

Sargon: If holistic is used then at some point it will become sequential and incremental--but that does not imply piecemeal, i.e., piecemeal not in the sense that you must read page one before page two, but piecemeal in the sense that page one is isolated from page 256. Piecemeal excludes whole-theme with a view toward becoming a whole-theme, but it may not. Like teaching a person a second language, distinctively. Is there a hidden variable that would account for the mentality or cross-over in the models. Seems like the distinctions between the following are important: intentional vs. unintentional, school vs. real-life, artificial vs. natural.

Lisa Baker: I believe the piecemeal approach versus the whole-theme approach are mutually exclusive and that the two are considered opposites in every sense. It would be very difficult to use the approaches together, because the assumptions of each approach do not coincide. To break concepts into separate entitles is totally different than treating all components as a whole, based on a multisource nature of learning. The difference between the two can be compared to black and white. Although it is difficult for many educators to comprehend because for them to understand this whole-theme approach would mean that they would have to have a complete change in their way of teaching--a reorganization of insights. *I would like to see the whole-theme approach to become more than a theory, but to be actually integrated in the entire educational system in order to fulfill the insights of individuals according to relevant real-life situations.*

Note 1: Students composed their solutions to the mutual exclusion problem after a class discussion of the two approaches using the DLT organizer. *Note 2*: All the names in this Table are self-chosen pseudonyms.

appropriateness. For example, the reader may have already noticed that this paper is itself an application of the whole-theme approach to writing articles of this type. Thus, one way for the reader to try to evaluate the degree of transfer-appropriateness of the whole-theme approach to complex problem-solving is to reflect on how successfully the present paper has incorporated it into its own structure. The reader can verify this by trying to match the ongoing development of the ideas in this paper against the structure of the DLT organizer in Figure 1.

A Whole-Theme Interpretation of Learning Misconceptions

Each of the global coherence rings in the cylinder-arrow diagonal of the DLT organizer is meant to be in a different color, representing a different theme with its own set of problems to be solved. A change from one theme to another (e.g., from red to blue in Figure 1) represents a radical reorganization of the learner's IKB. This is a change analogous to what Carey (1985) called strong conceptual change. In a thematic reorganization, an inference in the context of one theme might be viewed as a misconception in the context of the subsequent theme. Consider the question *Why do we pay for our food in a restaurant?* A child may respond because "we are hungry." As Carey suggests, a categorical structure such as a restaurant script cannot determine the boundaries for an inference like this because the same inference may occur in many other situations having nothing to do with restaurants. From the whole-theme perspective, inferences like this are far from being script-driven or schema-driven explanations. Rather, they represent a particular thematic organization of the person's IKB. Several years later, the same individual's answer may be an inference in the context of a radically different theme, where issues of the exchange of goods or services are involved in an economics course or discussion. An increase of a few dollars in the price of a meal may be judged as an insignificant price to pay by someone who is hungry. In the context of the economics theme, an increase of a few cents may be seen as having far-reaching consequences.

The whole-theme perspective is a theory of learning in specific domains. It is also a developmental theory. In this sense, a change from one to another global coherence ring in Figure 1 may represent a change from one developmental stage to another in a general, as opposed to domain-specific, developmental theory (e.g., a change from Piaget's stage of concrete operations to formal operations). In second language learning, it might represent a shift from one to another interlanguage (Selinker, 1972). Like the relatively domain-independent developmental or interlanguage stages, each of the domain-specific stages of IKB reorganization also generates its own characteristic misconceptions, although misconception might no longer be the most suitable term in the context of the whole-theme approach. This aspect of the whole-theme approach is compatible not only with the evidence from the literature on conceptual change (Carey, 1985) or developmental misconceptions (Caramazza, McCloskey, & Green, 1981; Rovegno, 1993), but also with background knowledge intrusions (Bartlett, 1932; Spiro, 1977; Steffensen, Joag-dev, & Anderson, 1979) and the development of scientific thinking (Gruber, 1989; Kuhn, 1962).

Whole-Theme versus Holistic

The term *whole-theme* as used here is consistent with, but not equivalent to, the term *holistic*. First, the whole-theme approach implies that the knowledge that represents the whole is thematic knowledge, which is qualitatively different from the categorical knowledge that represents the components of the whole (Iran-Nejad, 1989). This is an important qualification because it assumes that the whole can exist prior to and without the parts. In holism, the existence of the whole is dependent on the parts and not necessarily vice versa, even though the whole is more than the sum of its parts. In the whole-theme approach, the existence of the parts is dependent on the whole but not vice versa. This implication of the whole-theme approach is generally difficult for students to accept. The same students, on the other hand, have less difficulty finding a solution to the following problems: Explain how the parts can emerge out of the whole theme and *not* vice versa. Explain how the whole theme can exist without prior existence of the parts and *not* vice versa. One graduate student once used the analogy of the ocean and the waves to come up with solutions. She reasoned that the waves emerge out of the ocean but the ocean cannot properly be said to emerge out of waves and that the ocean exists prior to the waves and without their existence and not vice versa. This solution is strikingly similar in gist to the way the nervous system seems to create concepts out of thematic knowledge (Iran-Nejad, Marsh, & Clements, 1992).

Another fundamental difference between the whole-theme approach and holism is that, in the whole-theme approach, a clear distinction is made between whole-level (or theme-level) knowledge and unit-level or (concept-level) knowledge. In holism, the term *holistic* applies to the holistic aspect of concepts. This is the case also in the whole-theme approach in that concepts have a holistic aspect as they emerge out of thematic knowledge. By contrast, thematic wholeness, however, cannot be applied to the wholeness aspect of individual concepts. Rather, thematic wholeness is an aspect of a. entire domain of knowledge or the entire realm of the learner's IKB.

The notion of whole-theme, in the sense just described, suggests that the reorganizational influence of a new theme tends to permeate holistically throughout the entire realm of the learner's IKB. Thus, whole-theme implications tend to spread widely and deeply. If the learner believes that there is more to the universe than our solar system, the geocentric inference that the sun turns around the earth and the opposing heliocentric inference that the earth turns around the sun point to radically different whole-theme understandings whose respective realms expand far beyond the understanding of our immediate solar system; they tend to encompass the entire universe, just as the theme of a story tends to permeate throughout the entire story (Iran-Nejad, 1989), or even beyond the entire story in the form of the story moral. An example from Murphy and Medin (1985) may be used to illustrate this further. Given our intuitive knowledge of the world, it would be strange to believe that water is animate and still be able to make sense of phase relationships among water, ice, and steam. Now imagine that an extraordinarily well-prepared teacher could plant in us the seed to make us believe that water is indeed animate. Such an embryonic theme would tend to permeate, not just our concepts of water, ice, and steam, but the entire realm of our IKB, including whether or not life is possible on other planets.

And, finally, it is important to note that the focus of learning, as portrayed in the cylinder-arrow in the DLT organizer of Figure 1, is thematic knowledge, both as a process and an outcome. What makes learning possible, as a process, is the highly complex problem-solving context provided by the thematic organization. On the other hand, what is learned, as the outcome, is thematic knowledge itself. Individual skills, concepts, procedures, principles, propositions, vocabulary, and the like are (content-wise) units or pieces that can only exist and make sense in the context of the ongoing thematic organization and its evolution; without this context-wise influence, content-wise units would be as meaningless and as useless as Ebbinghaus' (1885/1964) nonsense syllables.

Summary and Implications

Simplification by Isolation

As teachers, when we think about introducing learners to a new domain, a justified sense of overwhelming complexity settles down on us along with a strong sense of urgency to simplify the task for our students. Concern for task simplification is perhaps the single most important influence on the organization of the traditional school curriculum. It is implicit in the hierarchical structure of educational taxonomies, where simpler learning objectives are arranged at lower levels of the hierarchy to be introduced and mastered first as prerequisites for more complex or higher-level objectives (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956; Woolever & Scott, 1988); it is an important practical consideration in determining the zone of proximal development (Vygotsky, 1978), a developmental bandwidth in which adult guidance is likely to be most effective; and it is a practical problem in optimal-level theory, which assumes an optimal level of interaction as a function of task difficulty or complexity (Berlyne, 1960; Hebb, 1955; Hunt, 1971; Iran-Nejad & Ortony, 1985).

The piecemeal cone-arrow in Figure 1 suggests that a domain gets simpler as we move in the opposite direction of the arrow. When teaching naive learners, the natural strategy would be to move all the way back to the beginning. With more advanced learners, we must find the knowledge level at which they are ready to receive instruction. More generally, this assumption of simplification by isolation has motivated researchers such as Ebbinghaus (1885/1964) to treat meaning as a confounding factor in the study of memory. Bartlett (1932) discussed this assumption and argued that it is not a necessary consequence of the experimental method of inquiry, as one might be inclined to think (see Iran-Nejad, McKeachie, & Berliner, 1990). Another manifestation of the assumption of simplification by isolation was behaviorism, which isolated complex tasks into sequences of observable stimulus-response connections. The assumption is alive and well three decades after the cognitive revolution and threatens to survive the widespread calls for reform (Wehlage, Smith, & Lipman, 1992).

A major problem with simplification by isolation is that it is accomplished directly at the expense of domain authenticity (Collins, Brown, & Newmann, 1989): As simplification moves further back toward the beginning of the cone-arrow in Figure 1, the essence of the complex domain evaporates with it as do critical thinking, thoughtfulness, and problem solving. Recent calls for reform, which propose an authentic curriculum as an alternative to the traditional curriculum, reflect a recognition of this problem. If this is true, at least two major obstacles must be removed before the authentic curriculum can become a reality. First, the assumption of simplification by isolation must be eradicated, roots and all. Second, alternative assumptions that protect domain authenticity must be identified. This paper examined both of these possibilities.

In the piecemeal approach, to simplify the learner's job, the teacher must solve the difficult problem of matching the learning task to the knowledge level of the learner. No one has ever found a reliable method of determining a student's background knowledge level. In schools, level determinations are made exclusively based on student grades, which take into account the learner's often very lean academic knowledge but almost never his or her rich IKB. As a result, most students learn to live in two very separate worlds: a school world from which they are continuously trying to escape and a real world that offers them no place to which to escape other than the school world. At the end, the story of schooling becomes the success story of the assumption of simplification by isolation.

Simplification by Reorganization and Integration

The whole-theme approach offers two principal ways for facilitating student learning, which can be practiced by teachers at all educational levels after a suitable training period. Teachers can help learners to reorganize their IKB globally by providing them with a DLT organizer that anchors and externalizes a new theme. This is the simplification by reorganization aspect of the process of teaching. The teacher provides the learners with a DLT organizer that might take the learners years, if ever, to discover on their own. Once established, such a DLT organizer can then be used by teachers to guide the process of simplification by integration, by posing the right kind of problems in the context of the DLT organizer to which learners can find their own answers. Integration in this sense, of course, refers specifically to DLT-IKB integration or the establishment of widespread continuity between individual student IKBs and their newly-established thematic knowledge.

The process of simplification by integration goes beyond the role played directly by the teacher. The DLT organizer can serve as a teaching map for the teacher, a learning map for students, a vehicle for posing and solving the problems that comprise the appropriate content of a course, and a means of communication among the teacher and learners as members of a learning community. Thus, implicit in the whole-theme approach are different assumptions about the nature of the course content, teaching, and learning.

The DLT organizer can facilitate learning or teaching by providing the learner or the teacher with a personal-growth learning or teaching map to use actively in her or his problem-solving or problem-posing explorations toward a PKB. Learning maps can help newcomers to a knowledge domain just as town maps can help newcomers to establish their personal territory in a new city. Therefore, thematic organizers provide a foundation for first-hand experience (as opposed to second-hand knowledge given to them by telling), autonomy, and self-reliance. For instance, if provided with Figure 1 early in a graduate course, students might use it to decide in the course of their domain-specific problem-solving (a) which of the two routes--portrayed by the two diagonals--they want to pursue, (b) what literature they want to consult, (c) what kind of knowledge--thematic or categorical--to make the focus of their learning activities, and (d) when they are ready to change their mind about a prior solution to a problem.

As scientists (like Copernicus and Galileo) have often had to discover, getting others to reorganize their own IKB may not be so easy a task when there is a firm commitment or a large investment by the learner in the older paradigm (e.g., the geocentric theory of the universe). For modern undergraduate, high school, or elementary school students, with no such commitment or investment, a willingness by the teacher to invest preparation time on the right kind of thematic organizer should go a long way toward helping the learner to gain a thematic grasp of the new perspective (e.g., heliocentric theory).

We have some preliminary evidence that learning as thematic reorganization encouraged by the DLT context followed by an adequate period of DLT-IKB integration might work exactly in the manner suggested by the whole-theme approach. Bea Volkman (Volkman & Iran-Nejad, 1993) used the DLT organizer of Figure 1 in her student teaching supervision project. Her goal was to use it as the context for helping student teachers reorganize the traditional school culture component of their IKBs, their knowledge of the university coursework, and their ongoing student teaching experience and to integrate these into a unified, authentic, whole-theme school culture. As the subjects moved through the supervision course, two aspects became evident. First, there was a tendency toward a shift from the traditional to a whole-theme school culture. Secondly, student teachers tended to understand and accept readily the thematic organizer; but they had a much harder time reintegrating and revaluating their traditional school culture assumptions, their university coursework, and their ongoing school experiences into this organizer.

Grasping a new theme is sometimes all that is necessary to put some learners on the right course of learning by problem-solving and discovery. These learners might then continue on their own, as opposed to waiting to be told by the teacher, to do extensive research to integrate their IKB into the newly discovered theme. Among these are those same kinds of learners that will go on to become self-made scientists and inventors in their respective fields. For other learners, the teacher may have to resort to the second principal way that the whole-theme approach offers for facilitating learning in students: by designing additional problem-solving activities, the teacher can offer learners opportunities to facilitate the integration of their IKB into the new theme (simplification by integration). It is important to note that simplification by integration must always follow simplification by reorganization and occur in the context of it. Without such a context, the same learning activities can readily change into simplification by isolation exercises. If we think of simplification by reorganization as being analogous to the teacher providing a map to a new city, simplification by integration would be roughly analogous to the teacher planning and conducting tours to major parts of the city and other similar activities.

The variety of thematic organizers and subsequent learning as problem-solving activities are limited by the imagination of the teacher and his or her willingness to invest time and other resources (Cochran, 1993). However, if the whole-theme approach is correct, imagination and willingness are necessary but not sufficient conditions for effectively facilitating student learning. Also essential is an indepth command of the content area as well as a rigorous program of preteaching and on-the-job training/experience with problem-solving in that domain. In other words, teachers must be fluent not only in their knowledge of the subject matter but also in the process of posing and solving problems that guide learners toward a PKB in that domain.

In the traditional culture of schooling, many practice education with the aid of a lean or no content-area knowledge and little more than a naive IKB. Therefore, it is not too far from the truth to claim that "public education--the industry in question--still uses the same methods it did a century ago . . . and [that] no other industry would last long with such a haphazard approach to self-improvement" (Marshall, 1993, p. 27). Earlier, I cited the data from the Raudenbush et al. (1993) study to show how the institutionalized culture of schooling represented in the cone-arrow in the DLT organizer of Figure 1 explains why teachers tend to resist setting critical thinking objectives for elementary and lower-rank students and postpone the teaching of critical thinking to more advanced and higher-rank students. A parallel argument applies to assignment of teachers to grades and to the amount of training and experience required of teachers. Specifically, if it is assumed that the teaching of the less advanced and lower-rank classes involves little more than presentation of basic concepts, facts, procedures, principles, and definitions, then a lean content area knowledge consisting of the same basic knowledge ought to be sufficient on the part of the teacher. The arguments in this paper suggest that what appears to be haphazard practice on the surface has deep root-level causes that make the traditional culture of schooling problematic.

The whole-theme approach suggests that all teachers must participate in a rigorous training program involving extensive preteaching and on-the-job problem solving. There are two problems with the naive IKB serving directly as a basis for making decisions about teaching, as is common in the traditional culture of schooling. First, the naive IKB represents the complex world Schon (1987) described as a slimy swamp of hard-to-manage problems. This is the complex end of the two-way complex-simple arrow on the right side of Figure 1. The simple end, relatively speaking, is the PKB which comprises a thematically organized body of domain-specific solutions gained after many years of domain-relevant problem solving. In other words, the learner's naive IKB is a good place for the teacher to begin the teaching of a complex domain; but the teacher's naive IKB is a dangerous foundation from which to teach. For the latter, the teacher must acquire and use a solid PKB, perhaps in the manner specified by the whole-theme approach. The second related problem is that the (intuitively-sound) solutions that the naive IKB produces are often irrelevant, or even counterproductive, to actual educational processes (see the irrelevant-relevant dimension in Figure 1). One such set of solutions, already discussed, is the one generated by the assumption of simplification by isolation.

Figure 1 portrays the distance between the naive IKB and expert PKB as an indefinite number of thematic reorganizations, requiring a rigorous problem-solving program of training, internship, and on-the-job practice. In fact, it has been recently suggested that teacher preparation must receive an emphasis similar in rigor and magnitude at least to that of preparing medical personnel such as surgeons and physicians (Iran-Nejad, Hidi, & Wittrock, 1992; Marshall, 1993). It is a reasonable conclusion, based on the arguments in this paper, that if educational research and training, as well as medical research and training, were to be conducted in the manner implied by the whole-theme approach, many of the problems that exist in today's education as well as in the area of health care would be resolved.

References

Anderson, R. C., Spiro, R. J., & Montague, W. E. (Eds.). (1977). *Schooling and the acquisition of knowledge*. Hillsdale, NJ: Erlbaum.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.

Bigler, P., & Lockard, K. (1992). *Failing grades*. Arlington, VA: Vandamere.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I, cognitive domain*. NY: Longman.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-42.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: The MIT Press.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition, 9*, 117-123.

Clancey, W. J. (1993). Situated action: A neuropsychological interpretation (Response to Vera and Simon). *Cognitive Science, 17*(1), 87-116.

Cochran, K. (1993). *Application of the whole-theme approach to teaching undergraduate educational psychology: Some personal observations*. Paper presented at the Twenty-second Annual Meeting of the Mid-South Educational Research Association, New Orleans.

Collins, A., Brown, J. S., & Newmann, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Hillsdale, NJ: Erlbaum.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.) New York: Dover Publications. (Original work published in 1885)

Greeno, J. G., & Moore, J. L. (1993). Situativity and symbols: Response to Vera and Simon. *Cognitive Science, 17*(1), 49-59.

Goodlad, J. L. (1984). *A place called school*. New York: McGraw-Hill.

Gruber, H. E. (1989). The evolving systems approach to creative work. In D. B. Wallace & H. E. Gruber (Eds.), *Creative people at work* (pp. 3-24). New York: Oxford University Press.

Hannaway, J. (1992). Higher order skills, job design, and incentives: An analysis and proposal. *American Educational Research Journal, 29*(1), 3-21.

Hebb, D. O. (1955). Drives and the C.N.S. (conceptual nervous system). *Psychological Review, 62*, 243-254.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace model. *Psychological Review, 93*, 411-428.

Hunt, J. McV. (1971). Intrinsic motivation: information and circumstance. In H. M. Schroder & P. Suedfeld (Eds.), *Personality theories and information processing* (pp. 85-130). New York: Ronald Press.

Iran-Nejad, A. (1989). A nonconnectionist schema theory of understanding surprise-ending stories. *Discourse Processes, 12*, 127-148.

Iran-Nejad, A. (1990). Active and dynamic self-regulation of learning processes. *Review of Educational Research, 60*, 573-602.

Iran-Nejad, A., Hidi, S., & Wittrock, M. C. (1992). Reconceptualizing relevance in education from a biological perspective. *Educational Psychologist, 27*, 407-414.

Iran-Nejad, A., Marsh, G. E., & Clements, A. C. (1992). The figure and the ground of constructive brain functioning: Beyond explicit memory processes. *Educational Psychologist, 7*(4), 473-492.

Iran-Nejad, A., McKeachie, W. J., & Berliner, D. C. (1990). The multisource nature of learning: An introduction. *Review of Educational Research, 60*, 509-515.

Iran-Nejad, A., & Ortony, A. (1984). A biofunctional model of distributed mental content, mental structures, awareness, and attention. *The Journal of Mind and Behavior, 5*, 173-210.

Iran-Nejad, A., & Ortony, A. (1985). Qualitative and quantitative sources of affect: How valence and unexpectedness relate to pleasantness and preference. *Basic and Applied Social Psychology, 6*, 257-278.

Jenkins, J. J. (1974). Remember that old theory of memory? Well, forget it. *American Psychologist, 29*, 785-795.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Ladwig, J. G., & King, M. B. (1992). Restructuring secondary social studies: The association of organizational features and classroom thoughtfulness. *American Educational Research Journal, 29*(4), 695-714.

Marsh, G. E., II, & Iran-Nejad, A. (1992). Intelligence: Beyond a monolithic concept. *Bulletin of the Psychonomic Society, 30*(4), 329-332.

Marshall, J. (1993, December). Why Johnny can't teach. *Reason*, 27-31.

Marton, F. (1988). Describing and improving learning. In R. R. Schmeck (Ed.), *Learning strategies and learning styles*, p. 53-82.

Metz, M. H. (1978). *Classrooms and corridors: the crisis of authority in desegregated secondary schools*. Berkeley: University of California Press.

Meyers, C. (1986). *Teaching students to think critically*. San Francisco: Jossey-Bass.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *The Psychological Review, 92*, 289-316.

Newmann, F. M. (1990). Higher-order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies, 22*(1), 41-56.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Page, R. N. (1990). The lower track curriculum in a college preparatory high school. *Curriculum Inquiry, 20*, 249-282.

Peterson, P. L. (1988). Teaching for higher-order thinking in mathematics: The challenge for the next decade. *Perspectives on research on effective mathematics teaching* (Vol. 1. pp. 2-26). Reston, VA: Lawrance Erlbaum Associates.

Powell, A. G., Farrar, A. E., & Cohen, D. K. (1985). *The shopping mall high school: Winners and losers in the educational marketplace*. Boston: Houghton Mifflin.

Prawat, R. S. (1989). Teaching for understanding: Three key attributes. *Teaching and Teacher Education, 5*, 315-328.

Prawat, R. S. (1991). The value of ideas: The immersion approach to the development of thinking. *Educational Researcher, 20*(2), 3-10.

Raudenbush, S. W., Rowan, B., & Cheong, Y F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal, 30*(3), 523-553.

Rovegno, I. (1993). Content-knowledge acquisition during undergraduate teacher education: Overcoming cultural templates and learning through practice. *American Educational Research Journal, 30*(3), 611-642.

Schoenfeld, A. H. (1987). *Cognitive science and mathematics education*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schon, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics. X*, 209-30.

Shuell, T. J. (1990). Phases of meaningful learning. *Review of Educational Research, 60*, 531-547.

Spiro, R. J. (1977). Remembering information from text: The "state of schema" approach. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 137-165). Hillsdale, NJ: Erlbaum.

Spiro, R. J., Bruce, B. C., & Brewer, W. F. (Eds.). (1980). *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education*. Hillsdale, NJ: Erlbaum.

Steffensen, M., Joag-dev, C., & Anderson, R. (1979). A cross-cultural perspective on reading comprehension. *Reading Research Quarterly, 15*, 10-29.

Volkman, B., & Iran-Nejad, A. (1993). *Applications of the whole-theme approach to a student supervision project about holistic and piecemeal approaches to teaching*. Paper presented at the Twenty-second Annual Meeting of the Mid-South Educational Research Association, New Orleans.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1989). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 217-249). Hillsdale, NJ: Erlbaum.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wehlage, G., Smith, G., & Lipman, P. (1992). Restructuring urban schools: The New Futures experience. *American Educational Research Journal, 29*, pp. 51-93.

Woolover, R. M., & Scott, K. P. (1988). *Active learning in social studies: Promoting cognitive and social growth.* Glenview, IL: Scott, Foresman.

Zheng, Z. & Iran-Nejad, A. (1993). *A thematic-systematic approach to teaching tests and measurements.* Paper presented at the Twenty-Second Annual Meeting of the Mid-South Educational Research Association, New Orleans.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

MSERA Membership/Renewal Form

(Please print or type)

NAME: ____________________

TITLE: ____________________

INSTITUTION: ____________________

MAILING ADDRESS: ____________________
____
____

PHONE: ____________________ FAX: ____________________

ELECTRONIC MAIL ADDRESS: ____________________ BITNET___ INTERNET___
OTHER___
MSERA MEMBERSHIP: New___ Renewal ___
ARE YOU A MEMBER OF AERA? Yes___ No___
WOULD YOU LIKE INFORMATION ON AERA MEMBERSHIP? Yes___ No___

DUES:	Professional	$10.00	________
	Student	$6.00	________

VOLUNTARY TAX DEDUCTIBLE CONTRIBUTION
TO MSERA FOUNDATION ________

TOTAL ________

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. Joan Butler
1101 Yorkshire Road
Starkville, MS 39759

Research in the Schools
Mid-South Educational Research Asssociation
and The University of Alabama
P. O. Box 970231
Tuscaloosa, AL 35487-0231

NON-PROFIT ORG.
U. S. Postage Paid
Tuscaloosa, AL
Permit No.

TM-ESA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and The University of Alabama.

Volume 1, Number 2 **Fall 1994**

IN MEMORIAM: Ralph W. Tyler 1902-1994 1

Do Funding Inequities Produce Educational Disparity? Research Issues in the Alabama Case 3
Steven M. Ross, Lana J. Smith, John Nunnery, Cordelia Douzenis, James E. McLean, and Landa L. Trentham

Student Self-Concept-As-Learner: Does Invitational Education Make a Difference? 15
Paula Helen Stanley and William Watson Purkey

Self-esteem and Achievement of At-risk Adolescent Black Males 23
D. Lynn Howerton, John M. Enger, and Charles R. Cobbs

Sequential-Simultaneous Profile Analysis of Korean Children's Performance on the Kaufman Assessment Battery for Children (K-ABC) 29
Soo-Back Moon, Chang-Jin Byun, James E. McLean, and Alan S. Kaufman

An Analysis of the Charles F. Kettering Climate Profile 37
William L. Johnson and Annabel M. Johnson

Students and the First Amendment: Has the Judicial Process Come Full Circle? 47
Donald F. DeMoulin

Metaphor Analysis: An Alternative Approach for Identifying Preservice Teachers' Orientations 53
Janet C. Richards and Joan P. Gipe

The Effects of Violations of Data Set Assumptions When Using the Oneway, Fixed-Effects Analysis of Variance and the One Concomitant Analysis of Covariance 61
Colleen Cook Johnson and Ernest A. Rakow

Effects of Item Parameters on Ability Estimation in Item Response Theory 77
Jwa K. Kim

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of innovative teaching strategies in research/measurement/statistics, descriptions of technology applications in the classroom, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, Office of Research and Service, The University of Alabama, P. O. Box 870231, Tuscaloosa, AL 35487-0231**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personnal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1994 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 1, Number 2 **Fall 1994**

EDITORS
James E. McLean and Alan S. Kaufman, *The University of Alabama*

PRODUCTION EDITOR
Margaret L. Glowacki, *The University of Alabama*

EDITORIAL ASSISTANT
Anna Williams, *The University of Alabama*

EDITORIAL BOARD
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Michèle Carlier, *University of Reims Champagne Ardenne (France)*
Sheldon B. Clark, *Oak Ridge Institute for Science and Education*
Michael Courtney, *Henry Clay High School (Lexington, KY)*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Kentucky*
Donald F. DeMoulin, *Western Kentucky University*
R. Tony Eichelberger, *University of Pittsburg*
Daniel Fasko, Jr., *Morehead State University*
Patrick Ferguson, *Arkansas Tech University*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Samuel Hinton, *Eastern Kentucky University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
Randy W. Kamphaus, *University of Georgia*
Jwa K. Kim, Middle *Tennessee State University*
Jimmy D. Lindsey, *Southern University and A & M College*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Peter Melchers, *University of Cologne (Germany)*
Claire Meljac, *Psychologue AU C.H.S. Sainte-Anne (France)*
Soo-Back Moon, *Hyosung Women's University (Korea)*
Arnold J. Moore, *Mississippi State University*
Thomas D. Oakland, *University of Texas*
William W. Purkey, *University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Clemson University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Arkansas State University*
Bruce Thompson, *Texas A & M University*
Anne G. Tishler, *The University of Montevallo*
Wayne J. Urban, *Georgia State University*

GRADUATE STUDENT EDITORIAL BOARD
Vicki Benson, *The University of Alabama*
Ann T. Georgian, *The University of Southern Mississippi*
Jin-Gyu Kim, *The University of Alabama*
Robert T. Marsousky, *University of South Alabama*
Jerry G. Mathews, *Mississippi State University*
Dawn Ossont, *Auburn University*
Malenna A. Sumrall, *The University of Alabama*
Carol J. Templeton, *Mississippi State University*
Joan K. West, *Mississippi State University*

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 1

IN MEMORIAM: Ralph W. Tyler 1902-1994

The inaugural issue of *Research in the Schools* led with a summary of Ralph Tyler's contributions to research in the schools and included an interview that took place shortly after his 91st birthday. For those of us who knew Ralph Tyler, it was unimaginable that we would face the balance of our careers without his counsel close at hand. Ralph Tyler passed away on February 18, 1994, shortly before his 92nd birthday. He remained professionally active his entire life, giving an interview for Phi Delta Kappa as late as August 1993. Until the last year of his life, he traveled from coast to coast almost weekly, attending professional meetings, speaking, giving advice, and teaching. His fee was a good meal and some lively conversation.

My greatest memory of Dr. Tyler was his ability to ask questions that got right to the crux of a matter. Sometimes these questions made you feel foolish because the answer illuminated a deficiency in your thinking. However, he never asked questions in a threatening or ridiculing way. He was challenging you to solve your own problem. It was impossible to speak with Dr. Tyler without examining your own views. This man, whose advice was sought by presidents and other heads of state, gave the same attention and consideration to issues presented by anyone, regardless of rank or station.

As indicated in the introduction to the *Research in the Schools* interview, Dr. Tyler's influence on educational research spanned 60 years. His national recognition began when he became the evaluator in 1934 of the Eight Year Study, a study cited in the literature to this day. His work on the Eight Year Study is often credited with broadening the focus of educators from testing to evaluation. As the University Examiner at the University of Chicago, he continued to expand his ideas of measurement and evaluation. His influence is evident in the *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain* and the *Taxonomy of Educational Objectives, Handbook II: Affective Domain*. The senior authors of both these classic books were former students of Dr. Tyler. In the 1960s and 1970s, Dr. Tyler was asked and assisted with improving the educational systems in the (former) Soviet Union, Israel, Ireland, Indonesia, and many other countries. Also during the 1970s, 1980s, and 1990s, he served as a volunteer faculty member at numerous universities across the nation, sharing his ideas and keen intellect with new generations of students. During all of his career, his clear and concise publications broadened his influence on the field.

While Dr. Tyler is gone, his influence lives on through his deeds, his students, his ideas, and his writings. I feel that educational research and the field of education are better places for his having been a part. In closing, I can't help but remember one of his favorite questions, "But how will it affect the students?"

James E. McLean
The University of Alabama

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 3-14

Do Funding Inequities Produce Educational Disparity? Research Issues in the Alabama Case

Steven M. Ross, Lana J. Smith, and John Nunnery
University of Memphis

Cordelia Douzenis
Georgia Southern University

James E. McLean
The University of Alabama

Landa L. Trentham
Auburn University

The present study was solicited as part of the plaintiffs' legal defense in the Harper v. Hunt (1993) educational disparity litigation in Alabama. The study entailed a systematic study of school resources and conditions, a teacher survey, and a principal interview, conducted at 45 schools in either high-stratum (wealthy) or low-stratum (poor) school districts. Findings from all data sources were consistent in showing disparities favoring high-stratum schools on an overwhelming proportion (about 84%) of the variables examined. The impact on the trial decision supporting the plaintiffs is discussed along with the issue of balancing research protocol with courtroom needs.

In the past few years, educational equity litigation challenging the distribution of state education financing has been successful in at least five states: Texas, Montana, Kentucky, New Jersey, and Tennessee (see reviews by Brannan & Minorini, 1991; Brown, 1991; Odden, 1992; Policy Information Center, 1991). In New Jersey and Kentucky especially, the courts were persuaded by abundant evidence of the failure of public education in the states' poorest communities. In a case in Maryland (*Hornbeck v. Somerset County Bd. of Educ.*, 1983), however, the court ruled that, despite disparities that may exist between districts, there is no requirement for fiscal equalization that goes beyond providing a basic education. A fundamental issue in this decision was the lack of concrete evidence indicating if and how "disparity" translates into tangible educational impacts.

The above cases suggest inconsistencies and limitations in the ways that educational disparity has been researched in previous studies. First, such studies are usually conceived as "wealth-based" challenges to inequities between richer and poorer districts. The primary data presented to establish disparity are dispensation figures specifying per capita expenditures on various material and personnel resources by area or district. Lacking is information concerning the kinds and quality of resources provided in terms of curricula, after-school programs, parent involvement, special education, and other factors. Second, previous studies, with few exceptions (Mattson, Pace, & Picton, 1986), have relied on historical records (namely, state or district data bases) or subjective reports by school personnel (e.g., teachers, principals, superintendents) to support the case for disparity (e.g., Slavin, 1991). Although these data appear to provide valid indicators of nominal disparity, they do not reflect actual conditions of schools within the districts examined in terms of effective acquisitions and usage of resources.

Steven M. Ross is a Professor of Educational Psychology and Research and Associate Director of Research for the Center for Research in Educational Policy at the University of Memphis. Lana J. Smith is a Professor of Instructional Curriculum and Leadership at the University of Memphis. John Nunnery is a Research Associate for the Center for Research in Educational Policy at the University of Memphis. James E. McLean is a University Research Professor and the Assistant Dean for Research and Service in the College of Education at The University of Alabama. Landa L. Trentham is a Professor of Educational Research at Auburn University. Cordelia Douzenis is an Assistant Professor of Educational Research at Georgia Southern University. The authors wish to thank Brenda Johnson from Memphis State University and Judy Giesen from The University of Alabama for their help in the data collection. Please address correspondence regarding the paper to Steven M. Ross, CEPR, University of Memphis, Memphis, TN 38152.

For example, it is certainly conceivable that a school receiving one-half the per capita funding of a similar school might create a comparable or even superior educational environment as reflected by the physical facility, instructional programs, and teacher quality. *Funding* disparities suggest but do not necessitate *educational* inequalities.

The purpose of this study is to describe an educational study and results performed in association with the *Harper v. Hunt* (1993) litigation in Alabama. The plaintiffs in this case were parents of children in poor school districts in the state whose basic claim was a disparity in the educational opportunities that their children received relative to children in wealthier districts. The educational study was solicited as part of the plaintiffs' legal defense. The authors, as principal investigators, were interested in developing an investigative design that would be more comprehensive and powerful than those used in previous cases.

First, it was desired to triangulate information sources by using multiple measures of related variables. The specific methods selected included a school environment observation study, teacher survey, principal interview, and a cross-validation followup of the school environment study. Second, in the case of the school environment component, we wanted to design a methodology that would systematize and objectify the collection of data regarding school conditions and resources from site visitations. By comparison, evidence collected in the Montana studies (Mattson et al., 1986) consisted of summary impressions from visits by the principal investigators to plaintiff schools and higher-revenue schools. Such methodology seems likely to increase possibilities for the contamination of the data from observer bias. Third, we wanted to identify and employ methodologies for analyzing data and presenting results that would maintain the integrity of valid research practices while yielding appropriate case-relevant evidence. Specific research questions addressed were:

1. Do funding disparities between school systems in Alabama translate into differential allocations of educational resources for schools?

2. To what extent do funding and/or resource disparities correlate with observed conditions at selected schools (climate, educational resources, teacher attitudes, instructional programs, etc.)?

3. Are results pertaining to the above questions consistent across multiple data sources?

Method

School System Sampling

On-site visits were planned to 48 schools in 16 school systems. The sample of school systems was selected as representing the 8 highest and 8 lowest systems in local revenue per average daily attendance (ADA), as reported by the state for 1989. Local revenue was used as the criterion due to perceived limitations of state and federal funding as meaningful indicators of between-system disparity. Specifically, state funding is distributed at a fairly constant level across systems, thus resulting in minimal variation. Federal funding is earmarked for compensatory and supplementary programs that are designed to address the special needs of systems that serve disadvantaged students. Such funding, aside from making up a relatively small proportion of a system's total revenue, is thus inversely related to system wealth. Local revenue, on the other hand, comprises approximately 40% of total revenue for wealthier systems and varies by $3,000 per ADA across systems, due mainly to the abilities of the local counties or cities to raise funds through property taxes and other means.

Within each of the 16 systems, an elementary school, a middle school, and a high school were selected for visits and observations. For this selection, it was necessary to decide what criteria would be most appropriate for the purposes of the study. Given the small number of systems concerned, a random process was considered risky in the sense that selections might not be truly representative of typical schools in the low- and high-revenue strata. We therefore reasoned that using a correlate of school success, such as student achievement, in the selection would provide a basis for eliminating outlier schools. That is, a school that performed typically for a district would be unlikely to have unusual characteristics.

Two alternative strategies using standardized achievement scores (Alabama Basic Competency Test and Stanford Achievement Test, depending on grade) were suggested. The first strategy was regarded as the most valid from a research standpoint, the second most useful from a litigation perspective. Specifically, in the first approach, the median scoring school at each of the three grade levels would be selected. In the second approach, the highest scoring school at each level would be selected in the high-stratum system schools, whereas the lowest scoring school at each level would be selected in low-stratum system schools. The purpose of the latter approach would be to maximize the comparison of environmental conditions by contrasting ostensibly successful wealthy schools and unsuccessful poor schools. Since both approaches (median and maximum contrast) were judged to have merit in view of the study's objectives (research and trial), a combination strategy was adopted as a compromise. It involved using the maximum contrast selection for the four wealthiest and four poorest systems, and the median approach for the remaining four high-stratum and four low-stratum systems.

Using the above strategies, the sample of 48 schools was selected. Comparison of the median- and maximum-contrast approaches actually showed very little difference due to the fact that, in many of the systems, there was only one school at each level. School systems were contacted by a state education official to secure permission for the site visits. All systems agreed to participate, with the exception of one high-stratum system. Consequently, the sample consisted of 15 systems (7 high and 8 low) and 45 schools.

Instrumentation and Procedure

The purpose of the site visits was to document the types of facilities and the level of resources available for teaching and learning in the identified schools. On the basis of previous studies of facilities/resources, a number of pertinent site characteristics were identified and incorporated into the data collection procedures. Other variables were also included on the basis of the experiences and expertise of the research team. The resulting data collection procedures included an observational survey of facilities and resources, an interview of the principal of each school, and a teacher survey.

The observational study (School Environment Study) required that pairs of trained observers make a systematic tour of the school facility and document conditions relative to safety and security, grounds and playing fields, general exterior characteristics (buildings, walks, drives, etc.), interior building conditions (offices, classrooms, labs, rest rooms, cafeteria, library, gymnasium, lighting, etc.), equipment (desks, media, physical education, computers, etc.), and other resources (books, science materials, etc.). Altogether, 236 variables were assessed pertaining to these categories. Some involved counting resources and recording the total (e.g., number of swings, number of football fields), others involved making qualitative judgments of the condition or sufficiency of resources using 3-point or 5-point scales (e.g., adequacy of lighting, condition of windows, appearance of the teachers' room, etc.), and others involved indicating the presence or absence of a resource by checking "yes" or "no" (e.g., whether or not there was soccer field, a swimming pool, etc.). Space was also provided for observers to take notes of their impressions.

Twenty randomly selected teachers at each school (or all teachers if there were 20 or fewer at a school) were asked to respond confidentially to a survey addressing such topics as the adequacy of resources and supplies for teaching and learning, quality of facilities, use of time required for non-instructional activities, qualifications of teachers, and availability of aides.

A third instrument provided questions for a 15-20 minute interview with the principal of the school. Questions concerned class sizes, availability of qualified teachers and substitutes, and numbers and types of specialized classes (e.g., drama, psychology, foreign language) and extra-curricular activities.

Observers and Training Procedures

Observers were recruited from two sites at which research team members were available for supervision--Auburn University and The University of Alabama. All (n = 17) were either education graduate students or junior education faculty selected from a pool of applicants. Selection criteria included knowledge and experience in data collection and research, availability during designated periods of time, quality of work in other areas, and perceived ability to work well with school personnel.

Prior to the collection of data for the actual study, procedures and instruments were field-tested by two of the team members in a sample of public schools in Memphis, Tennessee. The field test revealed a high degree of consistency in the observer ratings on nearly all variables. Revisions to clarify the operationalization of certain variables and to facilitate the recording of data were also made. On the basis of the findings, a final version of the instrument, final training procedures, and an observer handbook were developed.

Two training sessions, one at each Alabama university site, were held for the observers. All training was conducted by one of the researchers who participated in the Memphis field test. During training, participants were guided through the materials and procedures that would be used for data collection, were given specific definitions and examples, and participated in discussions and question-and-answer activities regarding the procedures.

Data Collection Activities and Reliability Analyses

Arrangements for visits to the selected systems were made by research team members working directly with the system superintendent. Each superintendent notified the principals in the selected schools that members of the research team would be contacting them directly to make arrangements for specific dates and times for site visits.

Observers were scheduled in pairs to visit each school. The rationale for this procedure was that (a) two individuals would feel more confident than would one about carrying out the data collection procedures, asking questions, and exploring the school; (b) reliability checks could be conducted by determining the consistency of independently made observations by pair members; (c) where questions arose about particular variables, the two individuals could discuss them and identify a mutually agreeable response; and (d) having two observers would decrease the time needed to complete an observation at a given school.

Before visiting a school, the observation team contacted the building principal to make specific arrangements for the visit and the distribution and collection of the teacher surveys. Once on site, the team would first interview the principal and any other appropriate personnel (e.g., maintenance staff, media specialist, guidance counselor). The team members then toured the school facility, completed the observation forms, and collected the teacher surveys.

As noted above, each observer was required to participate in a reliability check. This involved having each member of the pair complete a separate observation form if either had not been checked previously. Once the observation was completed, the two observers were to compare their responses, without changing any, and then record their consensual response on a third form. This consensual form was then used in the data analysis. The original forms were spot checked by the first author to determine whether there was reasonable consistency (there was in all cases), and were later used in a formal inter-observer reliability analysis, the results of which are reported in a later section.

Cross-Validation Component

An additional aspect of the study was followup on-site visits by three of the principal investigators. The purpose of the followup was twofold: (a) to cross-validate information collected by the observer teams, and (b) to observe exemplary and extreme contexts firsthand. Due to time constraints, 8 schools (6 low stratum and 2 high stratum) were visited. The selection of schools was based primarily on two factors: (a) geographical location to permit the largest number of schools to be visited within the available two-day time period; and (b) schools likely to represent "clear cases" of disparity in resources and conditions. Thus, for this component, the interest was more to observe firsthand the extent of likely disparities than to conduct a controlled comparison of norms for each stratum. The investigators toured each school for approximately one hour, talked with principals and/or other personnel, made notes, and took photographs.

Results

School Environment Instrument

An inter-observer reliability analysis was conducted in three ways depending upon the type of data collected. For data involving dichotomous choices (yes/no) or 3-point rating scales, the percentage of times the two observers independently made the identical response was computed. For dichotomous responses, the average was 97%, and for 3-point ratings the average was 93%. For 5-point rating scales, Pearson correlations were computed for the pair ratings. The correlation coefficients ranged from .80 to .97 except for one anomaly. These results indicate very high degrees of consistency in observer responses.

A total of 236 variables from the School Environment Instrument were analyzed. Descriptive analyses involved constructing summary tables using a 2 strata (low vs. high) x 3 education levels (elementary, middle, secondary) format. For interval (and ordinal rating scales of 3 or more points), stratum-level means were displayed; for dichotomous variables, the percentages of "yes" responses were displayed. For variables representing counts of the quantity of resources (e.g., number of library books), adjustments were made for school size by dividing the total quantity by the average daily attendance (ADA) for the school. This adjustment increased the low-stratum school means relative to the high-stratum means due to the smaller ADAs for the former.

For directly comparing low- and high-stratum schools, significance tests, consisting of chi-square tests of independence and analysis of variance, were conducted on the overall (all education levels combined) data for each variable. Given the large number of separate analyses, and the concomitant inflation of the family-wise Type I error rate, these results were used mainly for identifying patterns or trends rather than for proving particular variables to be valid discriminators. Space limitations preclude reporting the results for each variable. Rather, a summary of *interval* (rating scale) variables that showed significant stratum differences is provided in Table 1. Table 2 presents a comparable listing for nominal variables. Each table also shows variables associated with stratum effects that were less than .05. When viewed cumulatively, these directional findings reflect patterns that were conveyed as evidence at the trial.

Altogether, for the 236 comparisons, 204 (84%) directionally favored the high-stratum schools, 24 (10%) favored the low-stratum schools, and 8 (6%) were equal. A total of 113 comparisons (48%) yielded effects with probabilities < .05, with all but one of these favoring the high stratum group.

As can be seen from the listings in Tables 1 and 2, the high-stratum schools had better maintained and more attractive school grounds, better athletic/playground facilities, brighter lighting in classrooms and hallways,

Table 1
Scaled Environment Variables Associated
with Significant Stratum Effects

	Stratum		
Variable	Low	High	*t*
Exterior Conditions			
School Grounds			
Grounds maintained[a]	2.38	3.38	-3.75***
	.92	*.86*	
Grounds clean[a]	2.78	3.43	-2.49*
	1.00	*.68*	
Safety			
Safety threats[a]	2.88	1.62	4.08***
	1.23	*.74*	
Safe from traffic[a]	3.00	3.62	-2.30*
	.88	*.92*	
Walkways & driveways			
Walkways flood[c]	1.77	1.25	2.71**
	.69	*.55*	
Walkway condition[a]	2.45	3.52	-4.06***
	.93	*.81*	
No. parking spaces[d]	155.9	292.60	-2.06*
	118.3	*297.7*	
Parking lot condition[a]	2.54	3.38	-3.55***
	.83	*.74*	
Driveway condition[a]	2.75	3.43	-3.36**
	.79	*.51*	
Exterior building conditions			
Age of school bldg. (in years)	36.75	26.05	2.27*
	18.96	*11.13*	
Bldg. attractiveness[a]	2.25	3.62	4.19***
	1.22	*.92*	
Windows clean[a]	2.50	3.19	-2.55*
	.98	*.81*	
Window condition[a]	2.67	3.62	-3.57***
	.82	*.97*	
Broken windows[b]	1.57	1.15	3.04**
	.51	*.37*	
Interior Conditions			
General			
Floor condition[a]	2.21	3.81	-5.61***
	1.02	*.87*	
Fountains appearance[a]	2.67	3.43	-3.03**
	.87	*.81*	
Fountains condition[c]	2.17	2.76	-2.72**
	.76	*.49*	
Ceilings appearance[a]	2.58	3.57	-3.57***
	.83	*1.03*	
Lighting			
Qual./hall lighting[c]	1.83	2.29	-2.91**
	.56	*.46*	
Qual./classroom lighting[c]	2.08	2.33	-2.15*
	.28	*.48*	
Health Facilities			
First aid supplies[c]	1.25	1.85	-3.56***
	.44	*.67*	
Lunchroom			
Attractiveness[a]	2.96	3.48	-2.31*
	.81	*.68*	
Cleanliness[a]	3.08	3.62	-2.36*
	.83	*.67*	
Rest Room Conditions			
Overall condition[a]	1.92	3.33	-5.15***
	1.06	*.73*	
Sanitary napkins[a]	1.00	1.95	-3.19**
	.00	*1.47*	
Toilet paper available[a]	3.00	4.29	-3.41**
	1.53	*.85*	
Toilet seats[a]	4.63	4.95	-2.03*
	.71	*.22*	
Soap available[a]	1.33	3.33	-5.40***
	.92	*1.53*	
Exhaust fans working[c]	1.25	1.89	-3.36**
	.53	*.88*	
Odor level[c]	2.04	1.48	2.34*
	.69	*.47*	
Towel holders[a]	2.29	3.48	-2.60*
	1.40	*1.66*	
Towels available[a]	1.25	3.29	-4.74***
	.90	*1.87*	
Lighting quality[c]	1.63	2.48	-3.91***
	.49	*.42*	
Rest room: Overall quality[a]	1.83	3.29	-6.06***
	.82	*.78*	
Rest room: Appearance[a]	1.75	3.24	-6.14***
	.85	*.77*	
Playground/Athletic Fields			
Elementary only			
Age of equipment[c]	1.55	2.33	-2.31*
	.69	*.52*	
Condition of equipment[a]	1.90	3.67	-5.13***
	1.30	*.82*	
No. Sandboxes[d]	0.00	3.40	-2.26*
	0.00	*4.30*	

Table 1 (continued)

Variable	Stratum Low	Stratum High	*t*
All levels			
Basketball courts - condition[a]	1.85	3.09	-2.42*
	1.46	*1.14*	
Spectator stands - condition[a]	2.55	3.45	-2.43*
	.82	*.93*	
Baseball fields - condition[a]	2.00	3.19	-2.80**
	.93	*1.38*	
No. Tennis courts[d]	.20	2.20	-3.28**
	.90	*2.80*	
No. Player benches[d]	.60	3.00	-2.03*
	1.50	*5.40*	
Gymnasium			
Girls' Locker room - attrct.[a]	1.88	3.44	-3.84***
	1.26	*1.03*	
Boys' Locker room - attrct.[a]	1.53	3.20	-5.00***
	1.01	*.86*	
No. Boys' lockers[d]	114.00	316.30	-2.33*
	250.80	*226.50*	
P.E. equip - quantity[a]	1.96	3.76	-6.36***
	.95	*.94*	
P.E. equip. - quality[a]	1.96	3.81	-6.25***
	.86	*1.12*	
Gym - condition[a]	2.18	3.61	-3.72***
	1.30	*1.09*	
Gym - attractiveness[a]	2.09	3.61	-4.44***
	1.19	*.92*	
Library - Media Center			
Library			
Attractiveness[a]	2.67	3.76	-3.75***
	1.00	*.94*	
Spaciousness[a]	2.63	3.86	-3.51**
	.97	*1.01*	
Cleanliness[a]	2.96	3.86	-3.51**
	.81	*.91*	
Media Center			
No. VCR players[d]	7.20	14.40	-3.19**
	6.60	*8.60*	
No. VCR cameras[d]	.90	1.80	-2.28*
	1.20	*1.40*	
No. Carousel projectors[d]	2.30	5.90	-2.53*
	5.30	*3.80*	
Classrooms/Offices			
Administrative Offices			
No. Desk computers[d]	2.40	5.80	-3.62***
	3.00	*3.20*	
No. Phones[d]	6.80	10.60	-2.55***
	3.60	*6.20*	
Attractiveness[a]	2.33	3.90	-5.66***
	.96	*.85*	

Variable	Stratum Low	Stratum High	*t*
Science Labs			
No. Science labs[d]	1.20	2.50	-2.06*
	1.40	*2.50*	
Quant. science equip.[a]	1.85	3.62	-4.27***
	.80	*1.26*	
Science equip. - qual.[a]	2.15	3.54	-3.11**
	.98	*1.27*	
Teachers' lounge			
No. chairs[d]	10.10	15.60	-2.42*
	7.00	*7.00*	
Attractiveness[a]	2.13	3.62	-5.36***
	.72	*.92*	
Auditorium			
Attractiveness[a]	2.30	3.67	-3.79***
	1.11	*1.19*	
Regular classrooms			
Attractiveness[a]	2.42	3.48	-4.04***
	.93	*.81*	
Desks - condition[a]	2.71	3.76	-4.17***
	.86	*.83*	
Lockers/cubbies[c]	1.26	1.60	-2.02*
	.45	*.82*	
A/V screen[c]	1.92	2.57	-3.14**
	.65	*.51*	
Globe[c]	2.00	2.43	-2.42*
	.42	*.47*	
Map[c]	2.00	2.38	-3.76***
	.00	*.50*	
Locking cabinets[c]	2.08	2.38	-2.59**
	.72	*.59*	
Wall clock[c]	2.25	2.57	-2.07*
	.53	*.51*	
Adequate shelf space[a]	2.04	2.71	-2.19*
	1.00	*1.06*	
Encyclopedias[c]	1.46	1.81	-2.12*
	.51	*.60*	
File cabinets[c]	2.33	2.76	-2.59*
	.64	*.44*	
Textbooks - condition[a]	2.46	3.43	-3.95***
	.66	*.98*	
Textbooks - availb.[c]	2.58	3.00	-2.92**
	.50	*.32*	
Teacher desk - cond.[a]	2.13	3.14	-4.45***
	.80	*.73*	

Note: $^{*}p < .05$ $^{**}p < .04$ $^{***}p < .001$.
[a]5-point scale. [b]4-point scale. [c]3-point scale. [d]Per 1,000 students (average daily attendance). Low-stratum schools: $n = 24$. High-stratum schools: $n = 21$.

Table 2

Nominal Environment Variables Associated with Significant Stratum Effects: Percentage of Schools with Selected Features by Stratum

	Stratum		
Variable	Low	High	χ^2
Exterior Conditions			
Playground/Athletic Fields			
Asphalt play surf.	4.20	45.0	10.36**
Separate soccer field	4.3	30.0	5.17*
Running track	0.0	65.0	21.43***
Walkways/Driveways			
Crossing guard	12.5	63.2	11.98***
Entr./Exit signs	34.8	85.7	11.78***
Auto drop-off	70.8	95.2	4.56*
Interior Conditions			
General			
Student lockers	66.7	81.0	4.56*
Heating/cooling			
Central air	20.8	100.0	28.77***
Wall units (A/C)	91.3	30.0	17.21***
Communications			
PA system	83.3	100.0	3.84*
Student public phone	29.2	61.9	4.86*
Faculty phone	25.0	90.5	19.45***
Health facilities			
Bed available	8.3	42.9	7.23**
Library-Media Center			
A/V Production	16.7	50.0	5.59*
Gymnasium			
Soap/Boys' lockerroom	0.0	40.0	7.94**
Football equipment	87.5	100.0	7.25**
Tennis equipment	20.8	76.2	13.79***
Gymnastics equipment	29.2	71.4	8.00**
Soccer equipment	62.5	95.2	6.95**
Classrooms/Offices			
Administrative offices			
FAX machine	0.0	28.6	7.91**
Teachers' lounge			
Telephone	5.9	61.9	12.67***
Regular classrooms			
Exposed pipes	45.8	9.5	7.19**
Special classrooms			
Music			
Music room	20.8	90.5	21.83***
Band			
Band room	43.5	75.0	4.37*
Music stands	71.4	100.0	5.27*
Other special rooms			
Foreign language lab.	4.3	28.6	4.81*
Art room	0.0	100.0	44.00***
Home economics	54.2	61.9	12.67***
Auditorium			
Sound system	45.8	85.0	7.23**
Working microphone	58.3	90.0	5.52*
Stage lights	43.5	75.0	6.59*
General			
Student bookstore	4.8	40.0	7.42**

Note: $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$. Low-stratum schools: $n = 24$. High-stratum schools: $n = 21$.

cleaner and better equipped rest rooms, better and more physical education equipment, more attractive and spacious libraries, more media equipment, a greater quantity and variety of special classrooms (e.g., music, art, band), and better equipped and more attractive classrooms. The only variable on which the low-stratum schools surpassed the high-stratum schools was the quantity of wall air-conditioning units. (This outcome, however, represents an unfavorable finding for the low-stratum schools due to such units being noisy and outdated relative to the central air conditioning systems installed in 100% of the high-stratum schools.)

To provide the most liberal picture of where low-stratum schools might have had advantages, Table 3 lists the variables on which the low-stratum means were directionally higher than the high-stratum means. It should be noted that many of these variables represent tabulations of the quantity of resources per ADA. Interpretations of how many of these comparisons are biased by the smaller ADA at the low-stratum schools are given in the Discussion section.

Table 3
School Environment Variables Showing Directional Advantages for Low-Stratum over High-Stratum Schools

Variable	Stratum Low	High	*t*
Portable classrooms[a]	4.70	5.50	-0.23
	13.0	*9.0*	
No. Apple micro-comp.[a] in library-media center	3.3	1.7	0.93
	6.1	*2.7*	
No. spectator stands[a]	2.8	2.0	0.71
	4.2	*2.0*	
No. IBM microcomp.[a] in library-media ctr.	7.5	3.4	0.63
	26.9	*3.8*	
No. lunchroom seats[a]	469.0	400.2	1.08
	236.0	*177.3*	
Lunchroom condition[b]	3.42	3.33	0.39
	.78	*.66*	
No. full gyms[a]	1.9	1.60	0.67
	1.5	*1.1*	
No. boys' toilets	4.25	3.65	0.84
	2.29	*2.46*	
No. girls' toilets	5.21	5.19	0.03
	2.11	*1.86*	
No. library seats[a]	111.4	97.8	0.69
	81.8	*39.8*	
No. library holdings[a]	17583	16832	0.69
	16626	*6070*	
No. weekly subscript.[a]	6.5	5.7	0.39
	7.5	*6.7*	
No. copiers/admn. off[a]	3.4	2.7	1.32
	1.8	*2.2*	
Variable	Low	High	χ^2
Shop room	54.2	38.1	1.16
Copier in teach. room	56.3	38.1	1.21
Wall unit A/C	91.3	30.0	17.21***
Science lab gas jets	67.2	64.3	0.07
Plygrnd.-prot. mats	12.5	0.0	0.81
Elect. in science lab	100.0	92.3	1.04
Science lab sinks	100.0	92.9	0.96

Note: $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$. Variables in the left column and top portion of right column are interval/ratio variables. Column entries are stratum means. Variables in the right column below the header are nominal variables; column entries are percentages of schools or rooms within schools for which the items were determined to be present. [a]Per 1000 students (average daily attendance). [b]5-point scale. Low-stratum schools: $n = 24$. High-stratum schools: $n = 21$.

Teacher Survey

A total of 421 low-stratum teachers and 404 high-stratum teachers completed the survey, a response rate exceeding 95% in both cases. Of the 16 items on which comparisons were made, significant stratum differences ($p < .05$) were obtained on 11 (69%), with all (100%) directionally favoring the high-stratum schools. The significant variables are summarized in Table 4. Among the advantages indicated for the high-stratum schools are teacher perceptions of more adequate resources, better room conditions, more planning time, fewer demands for fund raising activities, and increased support for travel funds and teacher aides.

Principal Interview

The principal interview yielded data on 22 variables. Of these, 20 (91%) directionally favored the high-stratum schools. The exceptions were that low-stratum schools were more likely to have Channel One television (50% vs. 19%) and less likely to have combined grades (13% vs. 24%). Significant stratum differences were obtained on 9 (41%) of the variables (see Table 4). One variable was Channel One availability, while the others all favored the high-stratum schools, including smaller class size, number of teacher job applications, number of special classes (e.g., vocal music, foreign language, psychology), and the number of enrichment programs.

Discussion

The discussion of results will address two major areas: (a) findings from the research study, and (b) needs and decisions regarding the organization and presentation of the results for use as evidence for the *Harper v. Hunt* (1993) case. The latter issue addresses the problem of balancing scientific interests and ethical considerations with litigation needs.

The Research Findings

Findings from all data sources were consistent in showing clear disparities between the low-stratum and high-stratum schools. In fact, even though all four sources (environment study, teacher survey, principal interview, and visitation followup) directly examined many of the same or related variables, in no instance was a contradictory finding noted. Low-stratum schools were found to have less attractive physical plants and grounds, fewer educational resources in virtually all areas, fewer instructional offerings, and generally more dispirited staffs regarding their abilities

Table 4

Teacher Survey and Principal Interview Variables Showing Statistically Significant Stratum Effects

Variable	Stratum Low	Stratum High	t
Teacher Survey			
Adequacy of resourc.	2.41	1.68	13.49***
	0.86	*0.67*	
Classroom cool in hot weather	2.21	2.41	-2.36**
	1.14	*1.24*	
AC noise disruptive	2.65	2.43	2.43**
	1.30	*1.39*	
Teacher--fund-raising	2.27	1.92	7.38***
	0.62	*0.73*	
Student--fund-raising	2.35	2.17	4.04***
	0.62	*0.72*	
Avg. planning time	51.01	58.81	-5.24***
	17.45	*23.89*	
Extra pay--E/C Acts.	2.88	2.29	13.27***
	0.37	*0.80*	
Extra pay--intramurals	2.68	2.24	7.33***
	0.64	*0.90*	
Teach out of concentrat.	2.78	2.86	-2.59*
	0.43	*0.38*	

Variable	Stratum Low	Stratum High	χ^2
Teacher Survey (continued)			
Participate in F/R	67.6	42.4	51.90***
Travel funds avail.	22.5	81.6	272.01***
Teachers' aide (FT)	3.6	6.7	7.43**
Principal Survey			
Enrichmnt. programs	50.0	95.2	10.98***
Channel One	50.0	19.0	4.68*

Note: *p < .01, **p < .05, ***p < .001. Values in left column represent means and standard deviations by stratum; values in right column represent percentage responding "Yes." Low-stratum teachers: *n* = 421. High-stratum teachers: *n* = 404. Low-stratum principals: *n* = 24. High-stratum principals: *n* = 21.

to educate children effectively under existing conditions. The principal investigators found that, in every case (*n* = 45), they could read the observers' field notes "in the blind" and correctly guess from the descriptions whether the school was in the high- or low-stratum group.

Many of the discriminating variables listed in Tables 1 and 2 seem educationally important in the sense of giving children attending low-stratum schools disadvantages relative to their high-stratum counterparts. Examples included:

1. Restricted opportunities for participating in outdoor athletics such as soccer, basketball, and tennis.
2. Discomfort and distractions caused by noisy, antiquated, and inefficient heating and cooling equipment.
3. The negative ambiance of dark, old, and dirty school interiors.
4. The health risks and discomforts for children of having to use dirty, smelly rest rooms that often lacked toilet paper, soap, and towels. Where toilet paper was unavailable (in over half the rest rooms), the students were forced to bring their own or obtain it from a janitor.
5. Classrooms that lacked space, had unattractive and old furniture, and lacked learning resources such as textbooks for every child, globes, maps, encyclopedias, and projection screens.
6. Libraries that were old, unattractive, poorly stocked, and inadequately staffed.
7. Old (or no) gymnasiums with limited physical education equipment, deteriorating floors, and limited facilities and equipment.

Added to this list are the teacher and principal reports of staff and student involvement in fund-raising, lack of enrichment programs and special support subjects such as drama and foreign language, large class sizes, and limited funds to support professional development or to provide compensation for extra work. Clearly, teachers and staff at low-stratum schools work under conditions that are much more stressful and frustrating than is the case for their counterparts at wealthier schools.

At first glance, the results in Table 3 may appear to suggest advantages for the low-stratum schools on a fairly large group of variables. Consideration of the meaning of those findings, however, suggests otherwise. First, the only statistically significant effect showed a greater use in low-stratum of *wall* air-conditioning units, a negative condition compared with the newer, quieter, and better performing central units housed in all high-stratum schools.

Second, many of the directional advantages for the low-stratum were tabulations of the quantity of individual resources adjusted by ADA. Since ADA was lower at the low-stratum schools, this adjustment inflated the low-stratum mean for resources whose quantity would normally be invariant or insensitive to school size. For example, larger and smaller schools might both have one gym, similar weekly periodical subscriptions, and the same number of copiers in administrative offices. Thus, it seems of questionable importance that low-stratum schools had a greater ADA-adjusted quantity of seats in the library, weekly subscriptions, full gyms, copiers in offices and teachers' rooms, and auditorium seats.

Third, the greater quantity of computer resources in low-stratum schools is attributable to Chapter 1 funding for *supplementary* educational support. Since there was no reasonable way for observers to differentiate between Chapter 1 computers and computers acquired through the regular school budget, they were told to make an overall count of all computers and labs seen at the school. Even with the Chapter 1 acquisitions and the ADA adjustment, the differences between strata were relatively small and nonsignificant.

Fourth, the low-stratum advantages in three science lab resources (electricity, sinks, gas jets) are attributable to several of the high-stratum (but none of the low-stratum) *elementary* schools having science labs which were not so equipped, presumably for safety reasons. When the elementary schools are not included in the high-stratum averages, the advantages for the low-stratum schools are eliminated.

Fifth, the greater number of library holdings by the low-stratum schools seems attributable to two factors. One is the ADA adjustment noted above. The second is that, on the average, the low-stratum schools were 11 years older than the high-stratum schools, giving them considerably more time to acquire books. Not surprisingly, however, the books in the low-stratum schools were rated as older and in poorer condition than those at the high-stratum schools.

Sixth, the greater quantity of portable classrooms at the high stratum reflects not only the ADA-adjustment bias, but temporary conditions due to the rapid growth of schools in wealthier communities and new construction. These portable units tended to be new and in excellent condition compared to the older, seemingly permanent units at the low-stratum schools.

Seventh, the principal survey revealed a greater number of combined-grade classrooms at the high stratum. As with the portable classrooms, different causes for these conditions seem to prevail at the high and low strata. For high-stratum schools, such classes appear to be mainly a product of enrichment programs where younger middle school and high school students take classes, such as algebra and physics, with older students. At the low stratum, the main reason for combined grades appeared to be lack of classroom space and/or teaching staff.

Finally, the principal survey also indicated that significantly more low-stratum schools than high-stratum schools had Channel One television. This advantage seems largely due to the low-stratum schools' greater interest in acquiring the free television equipment that Whittle Communications' Educational Network provides to Channel One sites. Based on recent evaluation research by Johnston and Brzezinski (1992), the educational benefits of Channel One seem questionable.

Balancing Research Protocol with Courtroom Needs

The above research results provided what seemed to be compelling evidence of significant disparities in the educational opportunities available to children at high- and low-stratum schools. In preparing the results for presentation at the *Harper v. Hunt* (1993) trial, the principal investigators, as expert witnesses for the plaintiffs, developed a strategy of dissemination that combined rigorous research protocol with a more simplified presentation than the specialized professional field would require. Key dissemination strategies were as follows:

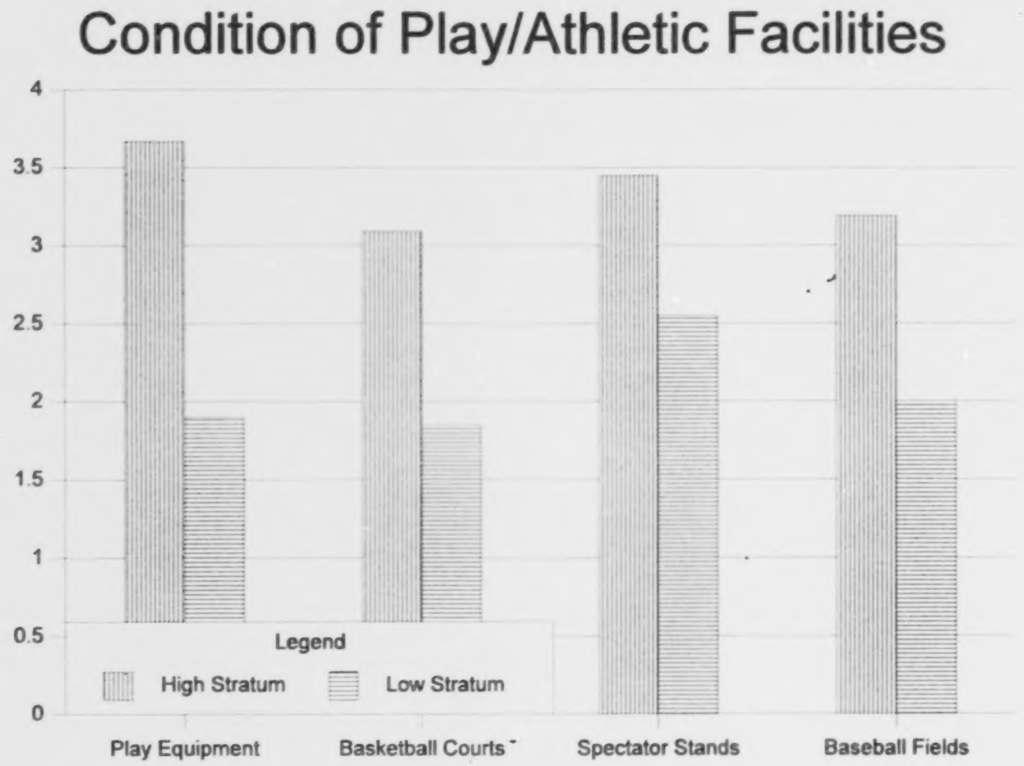
1. The key results presented were the proportion of directional differences favoring each stratum. We also included descriptive data on each variable from the school environment, teacher, and principal measures. The rationale was that the results represented a specific population of 45 schools in 15 systems.

2. We conveyed statistical results at the trial mainly through bar graphs showing stratum comparisons in bright contrasting colors. To illustrate, Figure 1 is a black-and-white overhead transparency similar to that presented during the first author's testimony.

3. We decided to give a high degree of emphasis to the presentation of photographs taken in the school visitation followup study. The combination of color bar graphs and photographs accentuating the associated stratum disparities was expected to be both attention-getting and memorable for the court. Where possible, photographs were mounted so as to pair low-stratum with high-stratum examples of the same resource (e.g., school entrances, playgrounds, libraries, etc.).

Conclusion

In presenting educational research results to the professional and scientific community, it is essential



Scale: 1 = below average, 3 = average, 3 = above average

Figure 1. Overhead similar to that shown at the trial: "Condition of play equipment and athletic fields.

that findings be accurate and valid. Appropriate attention must be given to factors such as validity threats, usage and outcomes of formal statistical analyses, and balanced presentations that give comparable coverage to positive and negative findings. In a courtroom presentation of research results, there is an ethical and legal commitment to present accurate information, but the mode of information dissemination must be adapted to a non-specialized audience.

In the Alabama study, combining scientific inquiry and trial objectives evoked relatively little strain, since the overwhelming positive evidence yielded through the former process was highly consistent with the latter. Based on our experiences, however, it is not difficult to imagine situations where the two domains would run directly counter to one another; i.e., the research findings support the opposing position. When such occurs, adherence to ethical values and sound scientific practices needs to prevail in the educational researcher's courtroom presentation of results.

On March 31, 1993, the Montgomery County Circuit Court ruled in favor of the plaintiffs. Judge Eugene

Reese, in a landmark decision, held that the Alabama Constitution guarantees all children in the state an adequate and equitable education, and that the state has not met that obligation. The decision ordered the state to provide "equitable and adequate educational opportunities . . . to all school children regardless of the wealth of the communities in which [they] reside" (*Harper v. Hunt*, 1993). The present study was cited numerous times in the text of that decision, with the associated testimony characterized by Judge Reese as "graphic and troubling." Reading the full document leaves little question about the significant impact of the study results on that decision.

References

Brannan, J. D., & Minorini, P. A. (1991, March). *Adverse impact on educational opportunity in cases challenging state school financing schemes.* (Commentary available from West Publishing Company, St. Paul, MN.)

Brown, S. I. (1991). Educational finance equity: Recent developments in state courts. *NASSP Bulletin*, 80-85.

Hornbeck v. Somerset County Bd. of Educ., 458A. 2d MD 758 (1983).

Harper v. Hunt (1993). CV-90-883-R combined with Alabama Coalition for Equity v. Hunt (1990). CV-90-883-4. Montgomery, AL: Circuit Court.

Johnston, J., & Brzezinski, E. (1992, April). *Taking the measure of Channel One: The first year.* Research summary. Ann Arbor, MI: Institute For Social Research, University of Michigan.

Mattson, R., Pace, M., & Picton, J. (1986). *Does money make a difference in the quality of education in the Montana schools?* Unpublished manuscript.

Odden, A. (1992). School finance in the 1990s. *Phi Delta Kappan*, 455-461.

Policy Information Center. (1991). *The state of inequality.* (Report, Educational Testing Service). Princeton, NJ: Author.

Slavin, R. E. (1991). *Funding inequities among Maryland school districts: What do they mean in practice?* Unpublished manuscript.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 15-22

Student Self-Concept-As-Learner: Does Invitational Education Make a Difference?

Paula Helen Stanley
Radford University

William Watson Purkey
University of North Carolina at Greensboro

The present study explored the relationship between invitational education and student self-concept-as-learner. The Self-Concept-As-Learner Scale (SCAL) was administered to 175 students in the seventh grade and readministered to the same 175 students in the ninth grade. During this period, invitational education was introduced and implemented throughout the school. Results indicated the SCAL scores of the students remained stable over the 2-year period. Self-concept-as-learner scores did not decline as predicted on the basis of the findings of previous studies.

The research described here is part of a larger study to determine the impact, if any, of "invitational education" (Purkey & Novak, 1984, 1988; Purkey & Schmidt, 1990; Purkey & Stanley, 1991) on student self-concept-as-learner. Invitational education is a theory of practice which maintains that every person and everything in and around schools add to, or subtract from, the process of realizing human potential. Ideally, the combined factors of people, places, policies, programs, and processes should be so intentionally inviting as to create a world in which each individual is cordially summoned to develop intellectually, socially, physically, psychologically, and morally.

Invitational education is centered on five propositions: (a) People are able, valuable, and responsible and should be treated accordingly; (b) education should be a cooperative activity; (c) process is as important as product; (d) people possess untapped potential in all areas of worthwhile human endeavor; and (e) potential can best be realized by places, policies, processes, and programs specifically designed to invite development, and by people who are intentionally inviting with themselves and others personally and professionally (Purkey & Novak, 1988).

In 1989, the first year of the present study, invitational education was introduced and implemented in a large (1,100 students) junior high school in North Carolina. Its implementation, which continued over a 2-year period, was made possible thanks to funding provided by an RJR Nabisco Next-Century-Schools Project. While the overall evaluation of the 3-year project is presently underway, this article addresses one facet of the investigation--changes in student self-concept-as-learner over a 2-year period.

Some Considerations Regarding Self-Concept

There continues to be a strong interest in the nature and function of self-concept in children and adolescents. However, studying self-concept can be a frustrating task. The hypothetical nature of self-concept seems confounded by confusion with respect to definition and assignment of causality as it relates to other variables, such as academic achievement or social competence (Kelly & Jordan, 1990).

In early studies, self-concept was defined by scientists as simply "self" (James, 1890). According to James, the self was composed of the material self, the social self, and the spiritual self. Cooley (1902), although recognizing that there were many "selves," focused on the social self. The social self, called the "looking glass self," is the result of recognizing and internalizing the evaluations of others. In other words, our perceptions of how others perceive us determines our self-concept. Many theorists believe that self-concept development is life-long and is learned from myriad experiences with the

---

Paula Helen Stanley is Assistant Professor of Counselor Education at Radford University and a Licensed Professional Counselor in private practice in Radford, VA. William Watson Purkey is Professor of Counselor Education at the University of North Carolina, Greensboro, and Co-Founder of the International Alliance for Invitational Education. Please address correspondence regarding the article to Paula Helen Stanley, Radford University, P. O. Box 6994, Radford, VA 24142.

external environment (Cooley, 1902; Harper & Purkey, 1993; Harter, 1986; Hattie, 1992).

A recent trend is the redefinition of self-concept as a composite of many dimensions, including cognitive, affective, and conative. Harter (1986) proposed that self-concept consists of domains that differ in significance for the individual according to one's age. Some domains are more significant at certain ages than others. For example, job performance, social competence, and appearance are components of self-concept which are salient factors in the definition of self . in adulthood. Scholastic competence, athletic competence, physical appearance, and peer acceptance are salient factors which define self in middle and late childhood.

Marsh (1993) developed a schema which divides self-concept into components, including academic self-concept and social self-concept. In addition, he studied math self-concept and school self-concept. Shavelson, Hubner, and Stanton (1976) developed a model of self-concept that is multidimensional and hierarchical in nature. This model is composed of academic and nonacademic components of self. Academic self-concept is comprised of self-concepts which relate to specific subjects. Nonacademic self-concept refers to social, emotional, and physical components of self.

Numerous studies have been undertaken to determine differences in self-concept among children as a function of gender, grade, race, and ability levels (Harper & Purkey, 1993; Hoge & Renzulli, 1993; Kelly & Jordan, 1990; Marsh, 1993; Winne, Woodlands, & Wong, 1982). Results have been mixed and often contradictory. For example, girls have been reported as having more negative self-concepts than boys (AAUW, 1991) and as having more positive self-concepts than boys (Harper & Purkey, 1993). Different findings may be attributable, at least in part, to the definition and treatment of self-concept. As recommended by Byrne, Shavelson, and Marsh (1992), future researchers might be advised to utilize instruments that give specific self-concept scores, such as self-concept-as-learner or social self-concept.

Instruments which have been used to measure self-concept-as learner or academic self-concept include the Self-Perception Profile for Adolescents (SPPA: Harter, 1986) and The Florida Key-Self-Concept-As-Learner Scale (SCAL: Purkey, Cage, & Fahey, 1973). The SPPA contains items which measure how one perceives one's academic ability, such as "I am smart," or "I can think of new ideas" (Harter, 1986). The SCAL measures students' perceptions of their own behavior on four dimensions: relating (basic trust in people), asserting (trust in one's own value), investing (trust in one's potential), and coping (trust in one's own academic ability). These behaviors have been judged by teachers as reflecting a positive and realistic self-concept-as-learner. Each of these dimensions contributes to students' self-concept-as-learner which itself is a part of the "global" or total self-concept of the individual. The SCAL, originally developed as a method for teachers to infer student self-concept-as-learner, has been revised to allow students to rank themselves on behavioral indicators of self-concept-as-learner.

The Significance of Student Self-Concept

Student self-concept may provide a measure which is useful in assessing factors related to such concerns as underachievement, lack of school attendance, and dropping out of school. Because student self-concept reflects students' perceptions of their abilities, feelings of belonging in school, and perceived relationships with teachers and other students, it may be useful in planning preventive and remedial interventions in the school setting. It also suggests strategies for creating a total school environment which better meets the needs of all students in terms of gender, grade, and race.

A positive relationship between self-concept and academic achievement has been demonstrated by numerous researchers (Darakjian, Michael, & Knapp-Lee, 1985; Hansford & Hattie, 1982; Harter, 1983). There is consistent agreement among researchers and theorists that there is a definite relationship between students' evaluations of self as learner and their level of academic achievement (Burns, 1982; Byrne, 1984; Chapman, 1988; Eshel & Klein, 1981; Johnson, 1981; Purkey, 1970, 1978; Purkey & Novak, 1984). Students who have more positive perceptions of themselves and their abilities are more persistent at school tasks (Chapman, 1988), while those who have poor self-concepts are more likely to give up when faced with difficult situations (Covington, 1984).

Beane (1991) conducted an extensive review of ways that schools can work to enhance student self-concept. In his view, there are good reasons why schools should be concerned with enhancing self-concept: (a) Enhancing self-concept "is a moral imperative for schools, especially in a time when other social institutions and agencies seem unwilling or unable to provide support and encouragement in the process of growing up" (p. 25); (b) there is evidence that shows a correlation between self-concept and such behaviors as "participating, completions, self-direction, and various types of achievement" (p. 25); and (c) self-concept has broader ramifications than the personal development of an individual. Self-concept goes beyond the idea of:

> coping with problems and into personal efficacy or power. Conditions like racism, sexism,

> poverty, and homelessness detract from human dignity and for that reason debilitate one of its central features, self-esteem. The resolution of these issues will depend less on rhetoric and more on action, but action is not likely unless people believe they can make a difference. When looked at this way, enhancing self-esteem helps build the personal and collective efficacy that helps us out of the morass of inequity that plagues us. (p. 26)

Changes in Student Self-Concept over Time

Studies of student self-concept indicate a downward trend in student self-concept as students progress through school (Griffore & Bianchi, 1984; Harper & Purkey, 1993; Silvernail, 1987). Marsh (1993) reported that academic self-concept dropped for both boys and girls from grades 4 through 7. There was a general linear downward trend in general and academic self-concept for boys and girls in grades 2 through 7 in a study reported by Burnett (1993).

Harper and Purkey (1993) researched differences in self-concept-as-learner (SCAL scores) among average and gifted boys and girls in grades 6 through 8 and found a downward trend in both inferred and professed self-concept-as-learner of both gifted and average students. SCAL scores were lower for seventh and eighth grade students than for sixth graders. There was a significant decline in scores for students at all three grade levels over a 5-month period from fall to spring.

Efforts to Enhance Student Self Concept

Beane (1991) suggests that many attempts to improve self-concept have fallen short. Traditionally, schools have used three approaches. One approach involves such activities as sensitivity training. For example, students might sit in a circle and talk about how much they like themselves and others for 15 minutes one day a week. Another approach is the self-concept programs or courses taught during the school day. A self-concept curriculum, which is commercially or locally prepared, is taught to students. Beane suggests there may be more than 350 programs now, with 30 programs used with more frequency than others. There is little in the research literature that documents the value of packaged programs in promoting positive and realistic student self-concept-as-learner.

A third approach to addressing self-concept is to consider the importance of the school environment as a total system in which a positive and realistic self-concept can be fostered. This is the approach of invitational education.

The Invitational Education Approach

The concepts and application of invitational education are based on the assumption that students' behavior and achievement are largely influenced by the ways they view themselves and the world in which they live (Purkey & Novak, 1984; Purkey & Schmidt, 1987; Purkey & Stanley, 1991). Invitational education is anchored in self-concept theory and the perceptual tradition, both of which are concerned with the inner world of the individual. Each proposes that perception guides behavior. To understand other human beings, it is necessary to understand their unique perceptions of the world and of themselves.

Invitational learning proposes that there are five areas that create the "chemistry" of the school and that impact on student self-concept-concept-as-learner: people, places, programs, policies, and processes. Each of these five areas has a vital influence on meeting the needs of students, encouraging independent thinking, modeling good social skills, generating inclusion, and dealing with conflict constructively.

In addition, practitioners of invitational education support a systemic view of human development and change. Individuals live in systems. Schools, families, and communities are systems in which students develop their sense of self, learn how to relate to others, and develop the knowledge and skills needed to function in society. When one applies invitational education, it involves examining the entire culture of a school and the community within which the student exists.

Method of The Study

The present study focused on one area of student self-concept, that is, self-concept-as-learner, and sought to determine differences, if any, in scores among students using the variables of gender, grade, and race. The present study used a longitudinal design to determine differences in student self-concept scores for junior high school students over a 2-year period. Seventh grade students were administered the professed version of the SCAL and the same students were retested with the same instrument in the ninth grade.

The researchers hypothesized that scores of students who were exposed to invitational education in their school would not decrease from the seventh to the ninth grade as expected on the basis of the findings of earlier research studies (Griffore & Bianchi, 1984; Harper & Purkey, 1993; Silvernail, 1987).

Participants

Participants in the study included 175 junior high school students who were administered the professed version of the SCAL while in the seventh grade. The SCAL was administered to the majority of the student body at the junior high school. However, there were only 175 students who reported self-concept scores for **both** the 1990-91 and 1992-93 school years and therefore comprised the sample for the longitudinal study. The same students who were administered the SCAL in the seventh grade were retested in the ninth grade. The smaller number of participants (N = 175) was due to school transfers, unusable tests, and student absences during the day of the test administration.

Other participants in the study were the faculty and staff of the junior high school who received training in invitational education. The entire faculty and staff (100) participated in two full-day training programs in invitational learning, while faculty team leaders participated in additional 2-day training sessions. The faculty who administered the SCAL to students also received training in the administration of the instrument.

Instrumentation

The SCAL (Purkey et al., 1973) is a 23-item behaviorally-anchored instrument which has both inferred (teacher completes instrument for each student) and professed (student completes instrument for self) versions. The present study used the professed version. To complete the professed version, students select one of six options in response to each of 23 questions. The options consist of a Likert-type scale: 0 = Never; 1 = Very Seldom; 2 = Once in A While; 3 = Occasionally; 4 = Fairly Often; and 5 = Very Often. Items include "I get along with other students," "I keep calm when things go wrong," "I join in school activities," and "I do my school work carefully."

The SCAL was originally developed by asking teachers to identify classroom behavioral characteristics of students believed to possess positive and realistic self-concept-as-learners. The instrument was created to provide teachers with a relatively simple way to infer self-concept-as-learner of their students. An index of reliability of .84 was obtained for the SCAL. Coefficients of reliability employing split-half procedures ranged from .62 to .92. A split-halves estimate of reliability of .93 was determined (Purkey et al., 1973).

Four factors comprise the SCAL. These factors are relating, asserting, investing, and coping.

Relating reflects a basic trust in people. The student who scores high on the relating dimension probably identifies closely with classmates, teachers, and the school. He or she thinks in terms of **my** school, **my** teacher, and **my** classmates, as opposed to **the** teacher, **that** school, and **those** kids. Being friendly comes easy for this student, and he or she is able to take a natural, spontaneous approach to school life. The student finds ways to express feelings of frustration, anger, and impatience without exploding at the slightest problem.

Asserting suggests a trust in one's own value. The student who scores high on this factor has not learned to be helpless. Rather, the student feels control over what happens to oneself in school. The student who scores high on asserting is willing to challenge authority to obtain a voice in what is happening in the classroom. There seems to be present in this student a learned process of affirmation: to claim one's integrity, to compel recognition.

Investing refers to a student's trust in his or her potential. The student who feels good about oneself as learner is more willing to risk failure or ridicule. A high score on investing suggests an interest in originality and a willingness to try something new. This person often volunteers in class, although sometimes good intentions backfire. By investing, a student feels a release of emotional feeling and expresses an attitude of excitement and wonder.

Coping indicates a trust in one's own academic ability. The student who scores well on coping is interested and involved in what happens in the classroom. Pride is taken in one's work, and attempts are made to obtain closure. A characteristic of the individual who scores well on coping is that he or she can reasonably handle the challenges and expectations of school.

The contention of the SCAL is that when students **relate** well to others in school, feel able to **assert** their thoughts and feelings, feel free to **invest** in classroom activities, and, **cope** with the academic challenges of school, they demonstrate a "good" self-concept-as-learner.

Procedures

Students were selected for the study by class membership. Classes were randomly selected and teachers administered the SCAL to their classes during May 1991. Those seventh grade students in 1991 who were still enrolled in the school in May 1993 were readministered the SCAL. Teachers administered the SCAL during the same general time period.

In 1990, and continuing thereafter over a 3-year period, faculty and staff of the junior high school selected for this study were exposed to intensive staff development on invitational education. These programs included:

Introduction to invitational education. At the beginning of the 3-year project, all faculty and staff at the junior high school participated in a 1-day inservice program on invitational education. The purpose was to "break the mold" of traditional thinking. All participants were introduced to the concepts of invitational education, which included: (a) basic assumptions (trust, respect, optimism, and intentionality); (b) four dimensions (being personally inviting with self, being personally inviting with others, being professionally inviting with self, and being professionally inviting with others); (c) foundations of invitational learning (the perceptual tradition and self-concept theory); (d) levels of functioning (intentionally disinviting, unintentionally disinviting, unintentionally inviting, and intentionally inviting); and (e) five areas (people, places, policies, programs, and processes).

Small group workshops. Consultants for each strand conducted four day-long workshops for their specific strand during the first year of the project. Workshop content focused on generating ideas for school improvement, learning the process of organizational change, and creating plans for action, all from an invitational education orientation.

Leadership training. In addition to ongoing small group workshops at the junior high school, five members of each strand plus the school principal, the two assistant principals, and two school counselors participated in a 2-day training session at the University of North Carolina at Greensboro. The "Five P Relay" (Purkey, 1991), a technique modified from the work of MacIver (1991), was taught to participants in each of five strands (people, policies, programs, places, and processes) and then employed as a major training component. The relay involved asking each strand to set five clearly defined "do-able" goals; circulating the goals of each strand to the other four strands, in turn, who identify possible obstacles and ways to overcome these obstacles; and returning the list of goals, obstacles, and ways to overcome obstacles to the original area strand. Each of the five strands then developed an action plan.

Inservice programs. Workshops at the junior high continued for the next 2 years and focused on the following topics: classroom discipline, cooperative learning, student evaluation, advisor/advisee programs, interdisciplinary teaming, multicultural learning and awareness, leadership, and working with at-risk students. All workshop content was presented within the context of invitational education.

Data Analysis

SCAL scores of students for 1991 and 1993 were compared. A *t*-test procedure was used to determine significant differences between means of the SCAL scores for the first year and third year of the project.

An analysis of covariance was used to determine changes in SCAL scores for the same sample of 175 students. Interaction effects among the initial variable (initial SCAL scores) and other variables (gender, grade, and ethnic group) were tested.

Results

Results of the *t*-test comparing SCAL means from year 1 to year 3 indicated that there were no significant differences between means for the total SCAL scores. This was also true for the four subscales of relating, asserting, coping, and investing. As Table 1 indicates, means for the total score and four subtests remained stable from year 1 to 3. Although there were slight increases or decreases in some subscales, these changes were not statistically significant.

Table 1
t Test Procedure for Florida Key Total Score and Subscales

Group	N	Mean	Standard Deviation	t
		Total Score		
1991	175	75.00	16.00	-.5525
1993	175	74.00	18.32	
		Relating		
1991	175	18.36	4.27	.4634
1993	175	19.00	4.00	
		Asserting		
1991	175	12.00	4.22	-.5362
1992	175	12.00	4.34	
		Coping		
1991	175	26.00	5.5	-.4514
1992	175	26.00	6.1	
		Investing		
1991	175	19.00	6.50	-.9504
1992	175	18.00	7.33	

In Figure 1, results from the present study are compared with those of Harper and Purkey (1993). Harper and Purkey administered the SCAL in December and then, again, in April during the same academic year for two grade levels. Student SCAL scores decreased significantly over the period of 5 months for seventh and eighth graders. Although the time frame differed in the Harper and Purkey study, there is evidence to suggest that self-concept-as-learner has a tendency to decline during middle and junior high school.

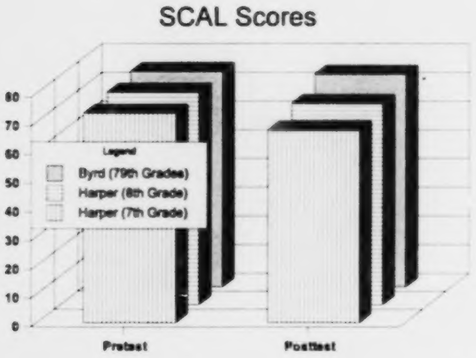


Figure 1. Mean pre- and posttest SCAL scores for the Harper and Purkey (1993) sample and the Byrd Junior High School sample. Mean pretest score for seventh graders in the Harper and Purkey study was 73. The posttest mean was 67. The pretest mean for the Byrd study was 75 and the posttest mean was 74.

Analysis of covariance was used to determine significant differences among variables for each subscale of SCAL. As Table 2 indicates, significant treatment effects were found for the variables of gender and race.

Results of earlier longitudinal studies using SCAL indicate that self-concept-as-learner scores drop as students move through grades 6 through 12. The results from the present study indicate a stability in scores rather than a drop as predicted and expected.

Table 2
Covariance Analysis for Total Scores

Gender	N	Mean (1991)	Mean (1993)	F	Pr>F
Female	91	77.13	77.13	2.48	.087
Male	84	73.00	71.00		
Race					
Black	59	72.30	69.40	3.03	.083
Other	41	73.00	73.00		
White	75	79.00	79.00		

One possible explanation of the stability in student scores is the implementation of invitational education. Teachers and staff participated on teams which identified weaknesses and strengths in five areas: people (interpersonal relationships); places (use of facilities); programs; policies; and processes (how people worked together to plan and solve problems). Over a 3-year period, changes occurred in the school which created a more positive school climate. School changes were recorded in five areas from year 1 to year 3 of the project:

People

1. Recognition of teachers increased from 10 activities in 1991 to 26 in 1993. Activities included being named "teacher of the week" and "teacher of the year," attending breakfast held in honor of faculty, and selection of faculty to attend special training in cooperative learning.
2. Parent/community volunteer hours increased from 1,344 in 1990 to 3,590 in 1993.

Places

1. The number of school beautification projects increased from 4 in 1990 to 8 in 1993.
2. Average daily circulation of library materials increased from 328 in 1990 to 365 in 1993.

Programs

1. Community/school partnerships increased from 15 in 1990 to 44 in 1993.
2. The number of dropouts decreased from 48 in 1990 to 14 in 1993.
3. Student scores on End of Course testing improved in 4 of the 6 areas tested.

Policies

1. The number of students retained at grade level decreased from 144 in 1990 to 110 in 1993.
2. The percentage of D and E grades decreased by 8%.

Processes

1. Total staff development hours attended by faculty increased by 30%.
2. Eight academic teams were in place in year 3, compared to 0 in year 1.

Discussion

Results of the present study suggest that the decline of student self-concept-as-learner over time can be ameliorated. The invitational education approach appeared to have had an influence on students' experience in school and thus their self-concept-as-learner.

To answer the question posed in the title of this study, it appears that invitational education did make a difference in student self-concept-as-learner. The results of the present study reveal a stabilization of student self-concept-as-learner scores rather than the decline noted in other studies that used the SCAL with similar groups of students. This was true for males and females and African-Americans and Caucasian students. The implementation of invitational education did coincide with stability of student self-concept-as-learner. Further research is needed to substantiate these findings. Replication studies may reinforce the results of this study: that invitational education can make a difference on student self-concept-as-learner.

The present study assumes that environmental changes initiated by the implementation of a model entitled invitational education within one school had significant effects on students' self-concept-as-learner over a 2-year period. Self-concept-as-learner scores appeared to remain stable rather than decline as reported in other studies. There may have been other factors, not yet identified, that could have influenced the stability of scores. For example, there were turnovers in faculty, with less than five faculty members leaving and less than five faculty members being reassigned to the school.

In future studies, the inclusion of a stronger comparison group in the research design would provide a stronger basis upon which to make assumptions concerning the present study. Without a stronger comparison group, the results of the present study must be interpreted cautiously.

References

American Association of University Women. (1991). *Shortchanging girls, shortchanging America: A call to action.* Washington, DC: Author.

Beane, J. A. (1991). Sorting out the self-esteem controversy. *Educational Leadership, 49*, 25-30.

Burnett, P. (1993, March). *Self-concept, self-esteem, and self-talk: Implications for counseling children.* Paper presented at the meeting of the American Counseling Association, Atlanta, GA.

Burns, R. (1982). *Self-concept development and education.* London: Holt, Rinehart, & Winston.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research, 54*, 427-456.

Byrne, B., Shavelson, R. J., & Marsh, H. W. (1992). Multigroup comparisons in self-concept research: Reexamining the assumption of equivalent structure and measurement. In T. M. Brinthaupt & R. P. Lipka (Eds.), *The self: Definitional and methodological issues* (pp. 172-203). Albany, NY: State University of New York Press.

Chapman, J. W. (1988). Learning disabled children's self-concept. *Review of Educational Research, 58*, 347-371.

Cooley, C. H. (1902). *Human nature and the social order.* New York: Scribner's.

Covington, M. V. (1984). The motive for self-worth. In R. Ames & C. Ames (Eds.), *Research in education: Student motivation* (pp. 77-113). Orlando, FL: Academic Press.

Darakjian, G. P., Michael, W. P., & Knapp-Lee, L. (1985). The long-term predictive validity of an academic self-concept measure relative to criterion of secondary school grades earned over eleven semesters. *Educational and Psychological Measurement, 45*, 397-400.

Eshel, Y., & Klein, Z. (1981). Development of academic self-concept of lower-class primary school children. *Journal of Educational Psychology, 73*, 287-293.

Griffore, R. J., & Bianchi, L. (1984). Effects of ordinal position on academic self-concept. *Psychological Reports, 55*, 263-268.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measure. *Review of Educational Research, 52*, 123-142.

Harper, K., & Purkey, W. W. (1993). Self-concept-as-learner of middle level students. *Research in Middle Level Education, 17*, 80-89.

Harter, S. (1983). Developmental perspectives on the self-system. In P. H. Mussen (Ed.), *Handbook of child psychology* (vol. 4, pp. 275-385). New York: Wiley.

Harter, S. (1986). Causes, correlates, and the functional role of global self-worth: A life-span perspective. In J. Kolligian & R. Sternberg (Eds.), *Perceptions of competence and incompetence across the life-span*. New Haven, CT: Yale University.

Hattie, J. (1992). *Self-concept*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoge, R. D., & Renzulli, J. S. (1993). Exploring the link between giftedness and self-concept. *Review of Educational Research, 63*, 449-465.

James, W. (1890). *Principles of psychology*. New York: Holt.

Johnson, D. S. (1981). Naturally acquired learned helplessness: The relationship of school failure to achievement behavior, attributions, and self-concept. *Journal of Educational Psychology, 73*, 174-180.

Kelly, K., & Jordan, L. (1990). Effects of academic achievement and gender on academic and social self-concept: A replication study. *Journal of Counseling and Development, 69*, 173-177.

MacIver, D. (1991, January). The "Pass it on" exercise presented at the Florida Regional Conference of the National Middle School Association. Fort Lauderdale, FL.

Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal, 30*, 841-860.

Purkey, W. W. (1970). *Self-concept and school achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Purkey, W. W. (1978). *Inviting school success: A self-concept approach to teaching learning*. Belmont, CA: Wadsworth.

Purkey, W. W. (1991). The 5-P relay: An exciting way to create an inviting school. *The Invitational Education Forum, 12*(2), 9-14.

Purkey, W. W., Cage, B. N., & Fahey, M. (1973). *The Florida Key manual*. University of North Carolina at Greensboro: International Alliance For Invitational Education.

Purkey, W. W., & Novak, J. M. (1984). *Inviting school success: A self-concept approach to teaching and learning*. Belmont, CA: Wadsworth.

Purkey, W. W., & Novak, J. (1988). *Education: By invitation only*. Bloomington, IN: Phi Delta Kappa Education Foundation Fastback 268.

Purkey, W. W., & Schmidt, J. J. (1987). *The inviting relationship*. Englewood Cliffs, NJ: Prentice-Hall.

Purkey, W. W., & Schmidt, J. J. (1990). *Invitational learning in counseling and development*. (ERIC Document Reproduction Service No. RI 88062011)

Purkey, W. W., & Stanley, P. H. (1991). *Invitational teaching, learning, and living*. Washington, DC: National Education Association.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*, 407-441.

Silvernail, D. (1987). *Developing positive student self-concept*. Washington, DC: National Education Association.

Winne, P. H., Woodlands, J. J., & Wong, B. (1982). Compatibility of self-concept among learning disabled, normal, and gifted students. *Journal of Learning Disabilities, 15*, 470-475.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 23-27

Self-esteem and Achievement of At-risk Adolescent Black Males

D. Lynn Howerton and John M. Enger
Arkansas Sate University

Charles R. Cobbs
Wynne Junior High School, Wynne, Arkansas

This pilot study investigated self-esteem and achievement of adolescent black males identified "at-risk" by their teachers. Forty-two junior high school students in a rural southern community were administered the Coopersmith Self-Esteem Inventory and the Stanford Achievement Tests. Self-esteem was significantly related to the achievement test battery composite ($r=.290$), and to science and mathematics subtests. Self-esteem was also related to average grade ($r=.426$), and to social studies and English grades. Relationships were also noted between self-esteem subscales and specific academic content areas.

Concerned black men in a rural, southern community developed a special program for at-risk black males (Cobbs & McCallum, 1992). Teachers were asked to identify at-risk participants in kindergarten through grade 8 for the program based on several characteristics (Cobbs, 1992). Two of these characteristics were low self-esteem and poor academic performance. The focus of this study was to investigate the relationship between self-esteem and academic achievement of the middle school students identified at-risk by the teachers.

Sewell (1985) noted there is little empirical evidence to show a relationship between school performance and self-esteem. More recently, Gaspard and Burnett (1991) observed a moderate relationship ($r = .38$) between self-esteem and grade average for rural ninth-grade students. This finding was similar to the results reported by Midkiff, Burke, Hunt, and Ellison (1986) for grade 8 students in a predominantly white suburb. However, Mboya (1986) found no significant relationship between self-esteem and standardized achievement test scores for 10th-grade black males. Similarly, Demo and Parker (1987) found no significant relationship between self-esteem and grade average for college-age black males. Kagan (1988) suggested further investigation is needed to examine the relationship between self-esteem and achievement in specific content areas.

Purpose of the Study

The purpose of this pilot study was to measure the relationship between self-esteem and academic achievement of at-risk adolescent black males. The Coopersmith Self-Esteem Inventory (SEI) was used to provide global measures of self-esteem (Coopersmith, 1967). School grades and scores from the Stanford Achievement Test (SAT) battery were used to measure academic achievement overall and in specific content areas.

Method

Forty-two black males in grades 6, 7, and 8 who had been identified at-risk served as subjects in this study. They ranged in age from 11 years 8 months to 16 years 7 months ($\bar{X}$ = 13.3 years, s = 1.09). A set of eight characteristics had been used by teachers to identify at-risk students for the Positive Impact Program for at-risk black males (Cobbs & McCallum, 1992). Along with the student's family status, the at-risk characteristics included low self-esteem, lack of motivation, poor academic record, chronic disciplinary problems, poor school attendance, poor hygiene and personal-care habits, poor social skills, and a disrespect for authority. A lower socioeconomic status was noted since three-fourths of the subjects received free lunch at school. Most of the subjects (88%) lived in a one-parent home or with a

---

D. Lynn Howerton is Professor of Psychology and Chair of the Department of Counselor Education and Psychology at Arkansas State University. John M. Enger is Professor of Education in the Department of Educational Administration and Secondary Education at Arkansas State University. Charles R. Cobbs is the Principal of Wynne Junior High School in Wynne, AR. Correspondence and requests for reprints should be addressed to Lynn Howerton, P. O. Box 1560, State University, AR 72467.

guardian. Two previous studies with these subjects have been reported (Enger, Howerton & Cobbs, in press; Howerton, Enger & Cobbs, 1993).

The school counselor administered the Coopersmith SEI to the subjects in one sitting. Prior to administering the SEI, release forms had been obtained from participants, parents, and school officials. Participants were requested to indicate "like me" or "unlike me" on an answer sheet, while being instructed that there was no right or wrong answer to each question. The SEI consists of 58 items and provides a global measure of self-esteem, ranging from zero to 100. The SEI also yields four subscales: general self, social self-peers, home-parents, and school-academic. SEI has been a popular research tool in assessing self-esteem and self-concept, having been cited in 942 articles over the past 25 years (Blascovich & Tomaka, 1991). In two reviews by the Buros Institute of Mental Measurements, SEI was recommended as an instrument appropriate for research (Peterson & Austin, 1985) and was noted for its wide applicability for research purposes (Sewell, 1985). The SEI is purported to be a reliable and stable instrument (Peterson & Austin, 1985; Sewell, 1985) with internal consistency reliability ranging from $r = .87$ to $r = .92$ for grades 4 to 8 (Coopersmith, 1981). Using an adapted SEI scale, Zirkel and Gable (1977) reported test-retest reliability of $r = .86$ for blacks.

For all students in grades 6, 7, and 8, SAT scores for the same year in reading, language, mathematics, science, social studies, and battery composite were recorded from school records. These SAT scores were converted to standard scores to reflect each student's relative standing in his class. School records yielded student grades in English, mathematics, science, and social studies for the school year. These grades were averaged to produce a grade average.

Results

Coopersmith Self-Esteem Inventory (SEI)

Applying coefficient alpha, item responses of the SEI for the 42 at-risk adolescent black males yielded an internal consistency reliability coefficient of $r = .793$. This reliability estimate of global self-esteem was somewhat lower than the measures reported by Coopersmith (1981) and Zirkel and Gable (1977). The reliability coefficients of the SEI sub-scales were: general self ($r = .678$), social self-peers ($r = .436$), home-parents ($r = .164$), and school-academic ($r = .458$). These findings support Zirkel and Gable's (1977) use of a modified SEI scale which omitted the home-parents subscale from the SEI global measure of self-esteem.

In this study, SEI was found to be a reliable measure of global self-esteem for at-risk adolescent black males. Moderate to strong internal consistency measures were found for three of the four SEI subscales: general self, social self-peers, and school-academic.

The SEI scores for the entire sample ranged from 38 to 96 with a mean of 63.0 and standard deviation of 12.75. Compared with normative data reported in the Coopersmith (1981) manual, this mean is similar to the mean of 64.6 for a sample of 60 black children in grades 3 to 8, but lower than the mean of 73.6 for 681 black children aged 8 to 14. The manual reported SEI means generally ranged from 70 to 80 with standard deviations of 11 to 13. For the distribution of scores in the present study, SEI scores at the 25th, 50th, and 75th percentiles were 54.3, 61.0 and 72.3, respectively.

The overall average self-esteem score for the at-risk middle school black males was significantly lower than most means reported in the normative studies in the Coopersmith (1981) manual. However, the average self-esteem score obtained in the present study was not significantly lower than means reported in studies for rural ninth graders (Gaspard & Burnett, 1991), for high school black males (Terrell, Terrell, & Taylor, 1988), and for blacks in grades 3 to 8 (Coopersmith, 1981).

Academic Achievement

Stanford Achievement Test (SAT). Students' overall and content area SAT scores were converted to standard scores (z-scores) representing their relative standing in their classes. These converted scores were determined by obtaining the overall and content area SAT means and standard deviations for all students in their middle school in grades 6, 7, and 8. The converted SAT scores for the at-risk black males had averages of: battery composite, $z = -.78$ ($s = 1.05$); reading, $z = -.79$ ($s = 1.08$); language, $z = -.67$ ($s = 1.00$); mathematics, $z = -.67$ ($s = .91$); science, $z = -.54$ ($s = 1.05$); and social studies, $z = -.74$ ($s = 1.02$). In summary, the average SAT scores for these at-risk black males generally fell .5 to .8 standard deviations below the mean of their middle school classes.

Grade averages. End-of-school grades in English, mathematics, science, and social studies for the at-risk black males averaged 1.85 ($s = .69$) on a 4-point scale (4 = A, 3 = B, 2 = C, 1 = D, 0 = F). Overall, these students had lower grades in science (GPA = 1.58, $s = .87$) than in social studies (GPA = 2.03, $s = .81$), English (GPA = 2.02, $s = 1.08$), and mathematics (GPA = 1.90, $s = .97$). Accumulating all of the subjects' grades across all four courses produced 3.1% A's, 21.4% B's, 39.0% C's, 30.8% D's, and 5.7% F's.

Relationship between SAT scores and school grades. The overall grade average correlated $r = .679$ with the SAT battery average. The correlations between grades and SAT scores in the four content areas were: English, $r = .380$; mathematics, $r = .447$; science, $r = .646$; and social studies, $r = .446$. These correlations were all significant at the .01 level.

Relationships Between Self-Esteem and Academic Achievement

SEI and SAT. As shown in Table 1, the SEI global measure of self-esteem was significantly related to the SAT battery composite ($r = .290$, $p < .05$), to SAT mathematics ($r = .308$, $p < .05$), and to SAT science ($r = .382$, $p < .01$). The SEI general self subscale was significantly related at the .05 level to only one of the SAT scores, SAT science. No significant relationships were identified between the SEI social self-peers subscale and SAT scores. The SEI home-parents subscale was significantly related at the .05 level to the SAT measures for battery composite, reading, language, mathematics, and science. For the SEI school-academic subscale, significant relationships were found at the .01 level with the SAT battery composite, reading, and language scores and at the .05 level with science scores. Of the five SEI measures (one global and four subscales), four were significantly correlated with SAT science; three with the SAT battery composite; two with SAT reading, SAT language, and SAT mathematics; and none with SAT social studies.

SEI and school grades. As shown in Table 2, the SEI global measure of self-esteem was significantly related to average school grades ($r = .426$, $p < .01$), English grades ($r = .309$, $p < .05$), and social studies grades ($r = .334$, $p < .05$). The SEI general self subscale was significantly correlated at the .05 level with average school grades and social studies grades. The SEI social self-peers subscale was significantly related at the .01 level to average school grades, English grades, mathematics grades, and science grades. The SEI home-parents subscale was significantly related at the .05 level to average school grades and science grades. The SEI school-academic subscale was significantly related only to English grades at the .05 level. Overall, of the five SEI measures, four significant relationships were found with average school grades, three with English grades, two with science grades and social studies grades, and one with mathematics grades.

Table 1
Correlations of Self-Esteem (SEI) and Academic Achievement (SAT) for At-risk Adolescent Black Males

	Self-esteem Inventory (SEI)				
SAT Scores	SEI Total	General Self	Social Self-Peers	Home-Parents	School-Academic
Battery Composite	.290	.182	.130	.325	.381
	(.035)	(.130)	(.212)	(.021)	(.008)
Reading	.252	.130	.061	.284	.458
	(.056)	(.209)	(.352)	(.036)	(.001)
Language	.236	.125	.052	.299	.399
	(.069)	(.218)	(.374)	(.029)	(.005)
Mathematics	.308	.242	.229	.279	.241
	(.027)	(.067)	(.078)	(.041)	(.067)
Science	.382	.298	.228	.358	.359
	(.007)	(.029)	(.076)	(.011)	(.011)
Social Studies	.189	.156	.014	.216	.232
	(.119)	(.166)	(.465)	(.088)	(.072)

p-values in parentheses

Table 2
Correlations of Self-Esteem (SEI) and Academic Achievement (Grades) for At-risk Adolescent Black Males

	Self-esteem Inventory (SEI)				
School Grades	SEI Total	General Self	Social Self-Peers	Home-Parents	School-Academic
Overall Average	.426	.307	.482	.315	.200
	(.003)	(.029)	(.001)	(.025)	(.111)
English	.309	.187	.401	.123	.269
	(.025)	(.121)	(.005)	(.222)	(.044)
Mathematics	.190	.161	.385	.072	-.085
	(.117)	(.158)	(.007)	(.328)	(.298)
Science	.233	.006	.394	.286	.229
	(.074)	(.485)	(.006)	(.037)	(.078)
Social Studies	.334	.268	.231	.230	.266
	(.019)	(.050)	(.078)	(.080)	(.051)

p-values in parentheses

Discussion

This pilot study found significant relationships between self-esteem and academic achievement for at-risk adolescent black males. The SEI global measure of self-esteem was significantly related to two composite measures of school performance, standardized test battery composite score ($r = .290, p < .05$) and end-of-year school grade average ($r = .426, p < .01$). The school grade relationship with self-esteem was similar to findings for rural ninth-grade students (Gaspard & Burnett, 1991) but not for college-age black males where no significant relationship had been reported earlier (Demo & Parker, 1987). Significant relationships between self-esteem and standardized test scores were found for at-risk adolescent black males in this study, which contrasts to the nonsignificant findings for 10th-grade black males (Mboya, 1986).

Kagan (1988) identified a need for the investigation of the relationship between self-esteem and achievement in specific academic content areas. In this study, significant relationships with self-esteem were noted for the four school performance areas investigated. The global measure of self-esteem was significantly related to standardized test scores in mathematics ($r = .308, p < .05$) and science ($r = .382, p < .01$), but no significant relationship was noted between self-esteem and reading, language, and social studies. Conversely, self-esteem was significantly related to school grades in English ($r = .309, p < .05$) and social studies ($r = .334, p < .05$), but not in mathematics and science. Thus, a significant relationship was found between global self-esteem and each of the four specific content areas, science and mathematics with SAT scores and social studies and English with school grades.

For these at-risk adolescent black males, several self-esteem subscales produced stronger relationships with specific academic content areas. The strongest relationships ($p < .01$) were noted between the school-academic subscale and standardized achievement measures in reading and language. The strongest SEI subscale relationships with school grades were noted between the social self-peers subscale with English, science, and mathematics grades.

As previously recommended in reviews of SEI (Peterson & Austin, 1985; Sewell, 1985), measures of self-esteem are useful for research and group interpretation. However, the user should be cautioned that SEI measures have limited utility in interpreting an individual's self-esteem.

The most appropriate application of these findings would be directed at programs working with at-risk black males, such as the aforementioned program. Programs such as PIP should be encouraged to include activities to enhance self-esteem and to improve academic performance of at-risk students. Continued efforts to boost self-esteem may serve to enhance students' academic performance; continued efforts to improve academic performance may increase students' self-esteem.

References

Blascovich, J. & Tomaka, J. (1991). Measures of self-esteem. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychology attitudes* (pp. 115-160). New York: Academic Press.

Cobbs, C. R. (1992). *The Positive Impact Program (PIP) for at-risk black males: An investigation of academic achievement and teacher ratings for adolescent black males at Wynne Junior High School.* Unpublished specialist's thesis, Arkansas State University, Jonesboro.

Cobbs, C. R., & McCallum, O. (1992, November). *Positive Impact Program (PIP) for at-risk black males.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Coopersmith, S. (1967). *The antecedents of self-esteem.* San Francisco: W. H. Freeman.

Coopersmith, S. (1981). *SEI: Self-Esteem Inventories.* Palo Alto, CA: Consulting Psychologists Press.

Demo, D. H., & Parker, K. D. (1987). Academic achievement and self-esteem among black and white college students. *Journal of Social Psychology, 127*, 345-355.

Enger, J. M., Howerton, D. L., & Cobbs, C. R. (in press). Internal/external locus of control, self-esteem, and parental verbal interaction of at-risk black male adolescents. *Journal of Social Psychology.*

Gaspard, M. R., & Burnett, M. F. (1991). The relationship between self-esteem and academic achievement of rural ninth grade students. *Journal of Rural and Small Schools, 4*, 2-9.

Howerton, D. L., Enger, J. M., & Cobbs, C. R. (1993). Locus of control and achievement for at-risk adolescent black males. *The High School Journal, 76*, 210-214.

Kagan, D. M. (1988). A discriminant analysis of alternative versus regular high school students. *The High School Journal, 71*, 60-68.

Midkiff, R. M., Burke, J. P., Hunt, J. P., & Ellison, G. C. (1986). Role of self-concept of academic attainment in achievement-related behaviors. *Psychological Reports, 58*, 151-159.

Mboya, M. M. (1986). Black adolescents: A descriptive study of their self-concepts and academic achievement. *Adolescence, 21*, 689-696.

Peterson, C., & Austin, J. T. (1985). Review of Coopersmith Self-Esteem Inventories. In J. V. Mitchell, Jr. (Ed.), *The ninth mental measurements yearbook* (pp. 396-397). Lincoln, NE: Buros Institute of Mental Measurements.

Sewell, T. E. (1985). Review of Coopersmith Self-Esteem Inventories. In J. V. Mitchell, Jr. (Ed.), *The ninth mental measurements yearbook* (pp. 397-398). Lincoln, NE: Buros Institute of Mental Measurements.

Terrell, F., Terrell, S., & Taylor, J. (1988). The self-concept of black adolescents with and without African names. *Psychology in the Schools, 25*, 65-70.

Zirkel, P., & Gable, R. K. (1977). The reliability and validity of various measures of self-concept among ethnically different adolescents. *Measurement and Evaluation in Guidance, 10*, 48-54.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 29-35

Sequential-Simultaneous Profile Analysis of Korean Children's Performance on the Kaufman Assessment Battery for Children (K-ABC)

Soo-Back Moon
Hyosung Women's University

Chang-Jin Byun
Kyungpook National University

James E. McLean
The University of Alabama

Alan S. Kaufman
The University of Alabama

The study investigated 440 Korean children, ages 2 1/2 - 12 1/2 years, tested on the Korean version of the Kaufman Assessment Battery for Children (K-ABC). The aim of the study was to determine whether Korean children demonstrated a profile on the K-ABC Sequential and Simultaneous Processing Scales characteristic of Japanese children. At each age, Korean children scored significantly higher on the Sequential Scale, in contrast to the High Simultaneous-Low Sequential profile displayed by Japanese children in previous investigations. Subtest analysis indicates an unusually strong ability for Koreans in Number Recall, relative to Americans. Implications of that finding for the documented high math ability of Korean children are explored. Also, the present results are contrary to Lynn's predictions regarding the intelligence of Oriental races.

The sequential-simultaneous dichotomy exists in research in diverse areas such as cognitive psychology, neuropsychology, and related disciplines (Kaufman & Kaufman, 1983a). Many researchers have dichotomized types of information processing: sequential versus parallel or serial versus multiple (Neisser, 1969), successive versus simultaneous (Das, Kirby, & Jarman, 1975; Luria, 1966), analytic versus gestalt/holistic (Levy, 1972), propositional versus appositional (Bogen, 1969), verbal versus imagery or sequential versus synchronous (Paivio, 1975, 1976), time-ordered versus time-independent (Bogen, 1975), and central versus automatic (Shiffrin & Schneider, 1977). There are many similarities in the definitions of serial, successive, and analytic/sequential processing; additionally, when scientists speak of parallel, simultaneous, or gestalt/holistic processing, it is evident that they are referring to a unified construct.

Sequential or successive processing refers to a person's ability to solve problems by mentally arranging input in sequential or serial order. Time and temporal relationships are important aspects of this type of processing since the stimuli tend not to be available at the same time (Kaufman & Kaufman, 1983a). The sequential processing of the stimuli is required regardless of the type of item content, method of presentation, or model of response. Simultaneous or holistic processing, on the other hand, refers to a person's ability to synthesize information in order to solve the problem. As with sequential processing, the content to be manipulated (e.g., semantic, figural, symbolic) is not as critical as the process.

Soo-Back Moon is an associate professor in the Department of Child Studies, Hyosung Women's University, Taegu, Korea. Chang-Jin Byun is a professor in the Department of Education at the Kyungpook National University, Taegu, Korea. Alan S. Kaufman is a Research Professor of Behavioral Studies in the College of Education at The University of Alabama and James E. McLean is a University Research Professor and the Assistant Dean for Research and Service in the College of Education at the University of Alabama. This paper is a modification of one presented by Soo-Back Moon at the 1988 Annual Meeting of the Mid-South Educational Research Association that won the Outstanding Dissertation Award. This paper was supported in part by Nondirected Research Fund, Korea Research Foundation. Please address correspondence regarding the paper to James E. McLean, Office of Research and Service, The University of Alabama, Tuscaloosa, AL 35487-0231 (Internet: JMCLEAN@UA1VM.UA.EDU).

Initial research evidence in support of the two information processing styles came from three sources (Kaufman & Kaufman, 1983a): (a) studies conducted by experimental and cognitive psychologists, primarily in a laboratory setting; (b) factor analytic work done by Das and his colleagues (Das, Kirby, & Jarman, 1979) in their pursuit of a partial validation of the Luria fronto-temporal versus occipital-parietal neuropsychological approach; and (c) experiments conducted primarily with split-brain patients (Springer & Deutsch, 1981). Subsequent investigations have demonstrated the robustness of the K-ABC sequential-simultaneous distinction for specific age groups throughout the preschool years (Kaufman & Kamphaus, 1984) and for separate groups of girls and boys (Kamphaus & Kaufman, 1986).

Analyses with the WISC-R produced readily identifiable sequential and simultaneous factors, with the former closely associated with the WISC-R Freedom from Distractibility factor and the latter with the Perceptual Organization factor (Kaufman & McLean, 1986, 1987; Keith & Novak, 1987; Naglieri & Jensen, 1987). The psychological meaningfulness of the sequential and simultaneous dimensions was further demonstrated by studies that investigated these constructs in concert with other theoretical constructs. Factors corresponding to this processing split emerged intact when the K-ABC was factor analyzed with fluid and crystallized scales derived from the Horn-Cattell theory of intelligence (Horn, 1989; Horn & Cattell, 1966, 1967), and when alternate measures of successive (sequential) and simultaneous processing were factor analyzed with measures of attention and planning ability in numerous investigations of Luria's (1980) neuropsychological theory (Das, Naglieri, & Kirby, 1994).

In Das et al.'s (1975, 1979) earlier work, successive and simultaneous factors consistently emerged, along with a speed factor, for a diversity of normal and exceptional samples, including culturally different groups such as white Canadian children and high-caste children from Orissa, India. Cross-cultural validation of the successive and simultaneous dimensions also has been provided for children from other locales in India and for Native Canadian, Australian, and Australian Aboriginal children (Das et al., 1994). And the proliferation of adapted and renormed K-ABC tests throughout the world has permitted cross-validation of the K-ABC sequential and simultaneous constructs in Europe and Asia. Clear-cut sequential and simultaneous factors emerged among samples of children from France (Kaufman & Kaufman, 1993), Germany (Kaufman, Kaufman, Melchers, and Preuß, 1991), and Japan (Matsubara, Fujita, Maekawa, Ishikuma, Kaufman, & Kaufman, 1994).

Additionally, Ishikuma, Moon, and Kaufman (1988) examined the simultaneous-sequential profile of Japanese children, ranging in age from 2 1/2 to 8 1/2 years, based on data provided by Lynn and Hampson (1986) on the Japanese version of the McCarthy scales. The results of that study offered evidence for a high simultaneous-low sequential profile for 2 1/2- to 8 1/2-year-old Japanese children. Kaufman, McLean, Ishikuma, and Moon (1989) used a similar methodology to examine Japanese children's relative performance in sequential and simultaneous processing from their scores on the Japanese WISC-R (Kodama, Shinagawa, & Motegi, 1978). Regression equations were derived from a sample of 170 normal American children who were tested on both the WISC-R and K-ABC in order to predict K-ABC sequential-simultaneous processing based on children's performance on WISC-R subtests. These equations were then applied to data obtained on Japanese children. The outcome of that study also supported a high simultaneous-low sequential profile for Japanese children; that hypothesized profile was not supported, however, in an investigation of Japanese children's performance on a Japanese translation of the K-ABC, using children from Alabama as a comparison group (Ishikuma, 1990).

The purpose of this study was to examine directly the simultaneous-sequential profile of Korean children for ages 2 1/2 through 12 1/2 years, to determine whether a high simultaneous-low sequential profile characterizes Asian children other than Japanese. Lynn (1987) has argued that Orientals, as a race, excel in visual-spatial (simultaneous) skills. Despite Ishikuma's (1990) negative finding in this regard, most investigations of Japanese intelligence support a high simultaneous-low sequential profile, in agreement with Lynn's (1987) theoretical position (Kaufman, 1990). The present investigation with Korean children offers a good opportunity to test out the generalizability of Lynn's theory, which he bases on evolutionary, empirical, and neurological factors.

The present study also provides an opportunity to understand a practical phenomenon: the fact that Korean children and adolescents easily surpassed the mathematical performance of children from a dozen nations, including the United States, in the International Assessment of Educational Progress (Lapointe, Mead, & Phillips, 1989). According to Kaufinan and Kaufman (1983a), good ability to process information sequentially "is closely related to a variety of school-related skills . . . [that] include memorization of number facts. . . . Sequential processing also may affect . . . applying the correct stepwise procedures for various mathematical skills such as 'borrowing'" (p. 30). It may be that the exceptional math ability by Korean children is partly a function of well-developed sequential processing.

Method

Subjects

The sample for this study was 440 Korean children who were tested between August and November, 1987, during the program for the development of the Korean version of the K-ABC. The sample was randomly selected from children ages 2 1/2 to 12 1/2 attending two elementary schools, four kindergartens, and five day care centers located in Taegu City, Korea. The sample was stratified at each age by sex and included 40 children in each of the 11 age groups, between 2 1/2 and 12 1/2 years. Equal numbers of boys and girls were included at each age. Socioeconomic status was not considered since there was no criterion available in Korea to determine the socioeconomic status.

Instrument

Moon (1988) translated the K-ABC Mental Processing subtests to Korean. The tryout version of the Korean K-ABC included direct translations of all items without modifications to the item content. Although a few of the pictures in the American K-ABC are not appropriate for Korean children (e.g., the "saw" in Magic Window and a "can opener" depicted in Matrix Analogies), Moon (1988) preferred to keep the K-ABC intact during the Korean tryout to permit the detection of all biased items by applying objective empirical techniques. The K-ABC Mental Processing Scales include three tests of Sequential Processing and seven tests of Simultaneous Processing. This 3:7 ratio does not represent the proportion of subtests actually administered to a given child. The ratio of simultaneous to sequential subtests is actually 2:3 at ages 2 1/2 and 3; 3:4 at ages 4 and 5; and 3:5 at ages 6 through 12 1/2.

Reliability data were analyzed to determine the consistency of a child's performance on the Korean version of the K-ABC Mental Processing Scales and subtests. Internal consistency reliability coefficients for the separate subtests for each scale were examined by using Cronbach's (1970) coefficient alpha. Mean reliability coefficients of Mental Processing subtests ranged from .79 (Gestalt Closure) to .87 (Triangles) for preschool children, whereas at the school age level the range was between .72 (Photo Series) and .79 (Triangles). Internal consistency reliability coefficients for the Global scales were computed using Guilford's (1954, p. 393) formula for determining the reliability of a composite. The mean coefficients ranged from .84 to .91 for preschool children. For school-age children, the mean coefficients were between .82 and .93, indicating good internal consistency at both the preschool and school-age level. Of particular interest for this study are the reliability coefficients for the Sequential and Simultaneous Processing Scales for Korean children. For preschool children, coefficient alphas averaged .89 for Sequential and .86 for Simultaneous; for age 5 and above, the mean values were .82 and .88, respectively.

Construct validity of the Korean version of the K-ABC Mental Processing Scales was also examined by principal factor analysis using 440 sample cases (Moon, 1988). Results of principal components analysis and principal factor analysis provide clear-cut empirical support for the existence of simultaneous and sequential dimensions.

Procedure

The Korean versions of the K-ABC Mental Processing Scales were administered by four trained graduate students in educational psychology. All children were tested individually in facilities provided by the school principals.

For each age group, raw scores on the Mental Processing subtests were computed. In order to facilitate comparisons with American children, the computed raw scores of each subtest were converted to scaled scores based on the American norms provided by the *K-ABC Administration and Scoring Manual* (Kaufman & Kaufman, 1983b). Then, sums of scaled scores for Sequential and Simultaneous Processing were computed by summing the designated scaled scores on the subtests; and the standard scores on these two processing scales, corresponding to the sums of the scaled scores, were identified from the norm tables in the Global Scales presented by Kaufman and Kaufman (1983b).

Mean standard scores on the Simultaneous and Sequential Processing Scales for Korean children were computed for each of the 11 age groups and the total sample of 440 children. Differences between mean standard scores on the two K-ABC processing scales were computed for the 11 age groups and for the total sample. The significance of these differences was then tested by two-tailed correlated *t* tests, applying the Bonferroni correction (Games, 1971) for multiple comparisons.

Results and Discussion

Table 1 presents the mean standard scores for Korean children on the Korean version of the K-ABC Sequential and Simultaneous Processing Scales, based on American norms. As indicated, Korean children have a distinct processing profile: high sequential-low simultaneous. The

Table 1
Mean Standard Scores for Korean Children on the K-ABC Simultaneous and Sequential Processing Scales

Age	N	Sequential Processing Standard Scores	Simultaneous Processing Standard Scores	Sequential minus Simultaneous Discrepancy	*t*
2 ½	40	128.4	108.2	20.2	7.19**
3 ½	40	117.7	95.4	22.3	6.22**
4 ½	40	111.8	99.3	12.5	7.23**
5 ½	40	122.4	109.4	13.0	6.55**
6 ½	40	112.6	105.8	6.8	3.08*
7 ½	40	119.4	108.3	11.1	5.50**
8 ½	40	124.6	111.2	13.4	7.33**
9 ½	40	120.7	104.5	16.2	9.00**
10 ½	40	121.7	110.9	10.8	5.98**
11 ½	40	122.8	110.9	11.9	5.54**
12 ½	40	128.6	115.2	13.4	7.07**
Total	440	121.0	107.2	13.8	19.48**

total sample of 440 children earned Sequential scores that were nearly one standard deviation (13.74 points) higher than their Simultaneous scores. The discrepancies at each age level and for the total sample of 440 proved to be statistically significant, favoring Sequential over Simultaneous processing ($p < .01$).

A high sequential-low simultaneous profile emerged consistently across all age groups. The sequential minus simultaneous discrepancies ranged from about 1/2 to 1 1/2 standard deviations and averaged almost 1 *SD* (13.8 points for the total sample). These findings are contrary to Japanese children's high simultaneous-low sequential profile (Ishikuma et al., 1988; Kaufman et al., 1989), and to Lynn's (1987) hypothesis that Oriental races as a whole excel in the kinds of visual-spatial tasks that compose the K-ABC Simultaneous Processing Scale as opposed to verbal-sequential tasks. In view of the Korean children's decisive difference (about one *SD*, on the average) in favor of Sequential Processing, relative to American children, the Lynn hypothesis--that posits a genetically programmed superiority in visual-spatial ability for Oriental nationalities--must be thoroughly reexamined. The results of this investigation cast great doubt about the generalizability of Lynn's (1987) hypothesis to Oriental nationalities, and Ishikuma's (1990) failure to detect a simultaneous superiority for the Japanese children he tested, relative to his American control group, suggests that Lynn's hypothesis may not apply unilaterally within Japanese samples. Further research on this important and provocative topic is needed.

Table 2 presents means and *SD*s for Korean children, relative to American norms, on each of the K-ABC Sequential and Simultaneous subtests. As a group, Korean children consistently performed relatively low on Face Recognition and Gestalt Closure, usually earning scaled scores in the 8-9 range. These subtests are the purest measures of what Horn (1989) calls Broad Visualization, a specific ability that requires "fluent" visual scanning, Gestalt closure, mind's-eye rotation of figures, and ability to see reversals" (p. 80). In contrast, Korean children demonstrated an astonishing short-term memory on Number Recall. Relative to children in the U.S., Korean children earned scaled scores of about 14 at ages 2 1/2 to 4 1/2; about 15 at ages 5 1/2 to 7 1/2; and about 17 at ages 8 1/2 to 12 1/2. The latter mean exceeds the American average by more than two standard deviations. From Horn's (1989) theory, the strength on Number Recall denotes excellent SAR, or Short-term Apprehension and Retrieval.

Stevenson, Stigler, Lee, and Lucker (1985) conducted a comprehensive study to examine whether children in three different cultures (America, Japan,

Table 2
Mean Standard Scores of the Korean Version of the K-ABC Mental Processing Subtests, by Age, for the Korean Sample (Based on American Norms)

Subtest		Sequential Processing			Simultaneous Processing						
Age		HM	NR	WO	MW	FR	GC	TR	MA	SM	PS
2 ½	Mean	13.8	14.8		12.8	9.7	10.9				
	SD	2.4	2.8		2.7	4.5	1.9				
3 ½	Mean	11.7	13.6		10.3	9.0	8.6				
	SD	3.8	2.8		3.9	2.5	2.8				
4 ½	Mean	11.0	14.2	10.4	10.3	9.5	8.9	11.4			
	SD	1.9	3.4	2.5	2.6	2.1	2.6	2.4			
5 ½	Mean	13.0	16.0	11.4			8.8	12.0	11.1	13.5	
	SD	2.3	2.1	1.7			3.2	2.4	2.4	2.3	
6 ½	Mean	10.5	14.2	11.2			8.5	13.0	10.9	11.8	10.2
	SD	2.8	2.9	2.7			2.0	2.7	2.8	2.6	1.4
7 ½	Mean	12.2	15.2	11.5			8.2	13.2	12.6	10.9	10.6
	SD	2.6	2.1	2.0			3.0	3.2	2.9	2.3	1.6
8 ½	Mean	12.0	17.0	12.4			8.7	14.4	12.6	11.5	11.7
	SD	2.8	1.7	2.0			2.3	1.6	2.3	2.3	2.4
9 ½	Mean	11.2	16.4	12.0			8.0	12.5	11.9	11.1	9.9
	SD	1.8	2.4	1.6			2.9	1.7	2.3	2.4	1.9
10 ½	Mean	10.3	17.4	12.4			8.9	14.0	12.0	11.8	11.0
	SD	2.4	1.5	2.2			2.6	2.2	3.0	2.1	2.0
11 ½	Mean	10.7	17.2	12.6			9.1	13.7	12.4	12.4	10.2
	SD	2.3	1.5	2.2			2.9	1.9	2.4	1.9	2.0
12 ½	Mean	12.0	17.1	14.0			9.7	14.2	12.6	13.1	11.4
	SD	2.2	1.1	1.5			3.0	1.4	2.2	2.1	1.5

Note. HM=Hand Movements; NR=Number Recall; WO=Word Order; MW=Magic Window; FR=Face Recognition; GC=Gestalt Closure; TR=Triangles; MA=Matrix Analogies; SM=Spatial Memory; PS=Photo Series

and China) differed significantly in their scores on cognitive tasks, including coding, spatial relations, perceptual speed, auditory memory, serial memory for words, serial memory for numbers, verbal-spatial representation, verbal memory, vocabulary, and general information. They found that the largest cultural difference occurred for serial memory of numbers, where Chinese children displayed remarkable superiority, but the superior serial memory of the Chinese children was not extended to words. The performance of Chinese children is in line with that of Korean children on the Korean version of the K-ABC Mental Processing Scales since their superior Number Recall did not extend to the same extent to Word Order and Hand Movements.

No hypotheses are apparent to account for the superiority of Korean children in the Number Recall subtest. Similarly, there are no easy explanations for their relative strength on the sequential processing of information in general, on tasks that require repetition of words and hand movements as well as numbers. Nonetheless, these strengths are substantial in nature and extend across the entire age range from preschool through elementary school; therefore, they cannot be accounted for simply by variables associated with schooling. The sequential strength is consistent with the findings that Korean 9- and 13-year olds far outstripped children from 11 other countries in their math abilities in the International Assessment of Educational Progress (Lapointe et al., 1989; Wainer, 1993). As noted previously, good sequential skills are needed for performing stepwise problems and memorization of number facts. Also, the whole mathematical system is a sequential, ordered system that lends itself to a sequential processing style.

Nonetheless, the conclusion that the Korean children's exceptional math performance is a result of their outstanding strength in sequential processing must remain speculative. For one thing, their superiority at age 13 was most evident in the area of **geometry** (Lapointe et al., 1989), a skill that seems more dependent on visual-spatial, **simultaneous** processes than on sequential syntheses. Also, the Korean children performed better than American children on all three Sequential subtests, but the superiority was demonstrated to a far greater extent on Number Recall (mean scaled scores of about 14-17) than on either Hand Movements or Word Order (means of about 11-13). Therefore, their math strength may reflect an unusual facility to manipulate numbers; in effect, then, their math strength may have influenced their exceptional ability to remember numbers rather than vice versa. Finally, other factors may have influenced their high math achievement even more so than their outstanding sequential processing, for example, higher motivation than children from most other countries (Wainer, 1993) or more time spent doing math homework than children from most other countries (Lapointe at al., 1989).

References

Bogen, J. E. (1969). The other side of the brain: Part I, II, and III. *Bulletin of the Los Angeles Neurological Society, 34*, 73-105; 135-162; 191-203.

Bogen, J. E. (1975). Some educational aspects of hemispheric specialization. *UCLA Educator, 17*, 24-32.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1975). Simultaneous and successive syntheses: An alternative model for cognitive abilities. *Psychological Bulletin, 82*, 87-103.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successive cognitive process.* New York: Academic Press.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence.* Boston: Allyn & Bacon.

Games, P. A. (1971). A multiple comparison of means. *American Educational Research Journal, 8*, 531-565.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 61-116). New York: W. H. Freeman.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253-270.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129.

Ishikuma, T. (1990). *A cross-cultural study of Japanese and American children's intelligence from a sequential-simultaneous perspective.* Unpublished doctoral dissertation, University of Alabama, Tuscaloosa.

Ishikuma, T., Moon, S. B., & Kaufman, A. S. (1988). Sequential-simultaneous analysis of Japanese children's performance on the Japanese McCarthy. *Perceptual and Motor Skills, 66*, 355-362.

Kamphaus, R. W., & Kaufman, A. S. (1986). Factor analysis of the Kaufman Assessment Battery for Children (K-ABC) for separate groups of boys and girls. *Journal of Clinical Child Psychology, 15*, 210-213.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1993). Joint exploratory factor analysis of the Kaufman Assessment Battery for Children and the Kaufman Adolescent and Adult Intelligence Test for 11- and 12-year olds. *Journal of Clinical Child Psychology, 22*, 355-364.

Kaufman, A. S., & Kamphaus, R. W. (1984). Factor analysis of the Kaufman Assessment Battery for Children (K-ABC) for ages 2 1/2 through 12 1/2 years. *Journal of Educational Psychology, 76*, 623-637.

Kaufman, A. S., & Kaufman, N. L. (1983a). *Interpretive manual for the Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1983b). *Administration and scoring manual for the Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *K-ABC manual d'interpretation.* Paris: Les Editions du Centre de Psychologie Applique.

Kaufman, A. S., Kaufman, N. L., Melchers, P., & Preuß, U. (1991). *K-ABC interpretationshandbuch.* Amsterdam: Swets & Zeitlinger.

Kaufman, A. S., & McLean, J. E. (1986). K-ABC/WISC-R factor analysis for a learning disabled population. *Journal of Learning Disabilities, 19,* 145-153.

Kaufman, A. S., & McLean, J. E. (1987). Joint factor analysis of the K-ABC and WISC-R for normal children. *Journal of School Psychology, 25,* 105-118.

Kaufman, A. S., McLean, J. E., Ishikuma, T., & Moon, S. B. (1989). Integration of the literature on the intelligence of Japanese children and analysis of the data from a sequential-simultaneous model. *School Psychology International, 10,* 173-183.

Keith, T. Z., & Novak, C. G. (1987). Joint factor structure of the WISC-R and K-ABC for referred school children. *Journal of Psychoeducational Assessment, 5,* 370-386.

Kodama, H., Shinagawa, F., & Motegi, M. (1978). *Manual for the Wechsler Intelligence Scale for Children--Revised* (standardized in Japan). Tokyo: Nikon Bonka Kagakusha.

Lapointe, A. E., Mead, N. A., & Phillips, G. W. (1989, January). *A world of differences: An international assessment of mathematics and science.* Report No. 19-CAEP.01. Princeton, NJ: Educational Testing Service.

Levy, J. (1972). Lateral specialization of the human brain: Behavior manifestation and possible evolutionary basis. In J. A. Kiger (Ed.), *Biology of behavior* (pp. 159-180). Corvallis: Oregon State University Press.

Luria, A. R. (1966). *Human brain and psychological processes.* New York: Harper & Row.

Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.

Lynn, R. (1987). The intelligence of the Mongoloids: A psychometric, evolutionary and neurological theory. *Personality and Individual Differences, 8,* 813-844.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan, and U.S.A. *Personality and Individual Differences, 7*(1), 23-32.

Matsubara, T., Fujita, K., Maekawa, H., Ishikuma, T., Kaufman, A. S., & Kaufman, N. L. (1994). *Interpretive manual for the Japanese K-ABC.* Tokyo: Maruzen Mates.

Moon, S. B. (1988). *A cross-cultural validity study of the Kaufman Assessment Battery for Children.* Unpublished doctoral dissertation, The University of Alabama, Tuscaloosa.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11,* 21-43.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Paivio, A. (1975). Imagery and synchronic thinking. *Canadian Psychological Review, 16,* 143-163.

Paivio, A. (1976). Concerning dual-coding and simultaneous-successive processing. *Canadian Psychological Review, 17,* 69-71.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Bulletin, 84,* 127-190.

Springer, S. P., & Deutsch, G. (1981). *Left brain right brain.* San Francisco: Freeman.

Stevenson, H. W., Stigler, J. W., Lee, S. Y., & Lucker, G. W. (1985). Cognitive performance and academic achievement of Japanese, Chinese, and American children. *Child Development, 56,* 718-734.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30,* 1-21.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 37-46

An Analysis of the Charles F. Kettering Climate Profile

William L. Johnson and Annabel M. Johnson
Ambassador University

This study focused on a multivariate analysis of the Charles F. Kettering School Climate Profile, a popular measure that is widely used to gather data for organizational planning and curriculum development. A total of 1,311 administrators, teachers, staff and students from an elementary, a junior high, and two high school campuses in a large school district in the Southwestern United States completed the General Climate Profile. Primary and second-order principal components analysis suggested different subscales from those given for the Kettering instrument. Subscale modifications are suggested to improve overall scale validity.

There is little question that organizational research occupies a popular position in the educational, psychological, industrial, and sociological literature. However, despite three decades of substantial empirical investigation, the meaning of organizational climate remains elusive (Anderson, 1982; Drexler, 1977; Guion, 1973; Halpin, 1966; Miskel & Ogawa, 1988; Moos, 1974; Stern, 1970; Tagiuri, 1968; Victor & Cullen, 1987). More recently, research on school effectiveness has generated a renewed emphasis on the importance of the educational environment in which optimal teaching and learning occurs (Good & Brophy, 1986).

Because of the conceptually complex and vague definitions of climate, James and Jones (1974) reviewed the major conceptualizations, definitions, and measurement approaches regarding organizational climate. Their popular and often-cited review was organized into three separate but not mutually exclusive approaches to defining and measuring organizational climate: (a) the multiple measurement-organizational attribute approach, (b) the perceptual measurement-organizational attribute approach, and (c) the perceptual measurement-individual attribute approach.

Representative of the multiple measurement-organizational approach is the definition of Forehand and Gilmer (1964) in which organizational climate is defined as a "set of characteristics that describe an organization and that (a) distinguish the organization from other organizations, (b) are relatively enduring over time, and (c) influence the behavior of people in the organization" (p. 362).

James and Jones (1974) also reviewed the perceptual measurement-organizational attribute approach (Campbell, Dunnette, Lawler, & Weick, 1970), which identifies four general categories of the organizational situation: (a) structural properties, (b) environmental characteristics, (c) organizational climate, and (d) formal rule characteristics. There is the possibility, though, that this approach may be inconsistent. In one sense, it proposed to measure organizational attributes that have been demonstrated to vary across levels of explanation such as the total organization, subsystem and group, whereas in another sense it is considered a psychological process that operates at a level distinct from objective organizational characteristics and organizational processes.

James and Jones (1974) addressed the perceptual measurement-individual attribute approach, which characterizes organizational climate as an individual's set of general or global perceptions about his or her organizational environment. These general perceptions reflect the interaction between personal and organizational characteristics in which the individual forms perceptions about overall climate.

Many of the characteristics of the perceptual measurement-individual attribute approach to climate research were identified in work by Schneider and his associates (Schneider, 1972, 1973; Schneider & Bartlett, 1968, 1970; Schneider & Hall, 1972). Schneider and Hall described organizational climate as

---

William L. Johnson is a Professor, Chair of the Psychology and Education Department, and Associate Dean of Academic Affairs at Ambassador University in Big Sandy, Texas. Annabel M. Johnson is a Professor of Home Economics at Ambassador University in Big Sandy, Texas. Her specialty is in the area of management. Please address correspondence regarding the paper to William L. Johnson, Associate Dean of Academic Affairs, Ambassador University, Big Sandy, TX 75755.

a set of global or general perceptions held by individuals about their organizational environment. Like James and Jones (1974), they wrote that summative perceptions mirror the interaction between personal and organizational characteristics and that individuals form their perceptions about the overall organizational climate from this interaction. Climate is viewed as an individual perception and is described as personalistic. It is seen as a general perception or intervening variable based upon the interaction between the individual and the environment.

Although this approach assumed that situational and individual characteristics interact to produce a third set of perceptual intervening variables, such an assumption does not mean that perceived climate is not an individual attribute. Rather, the intervening variables are individual attributes that provide a link between the situation and the behavior. Climate is treated as an individual attribute because it is the individual's perceptions that are important, not the objective situation (Guion, 1973). Such an approach appears to provide a step toward the formulation of specific theoretical statements regarding the nature of the psychological process between the organizational situation and the attitudes and behavior of individual members of the organization (James & Jones, 1974).

During the 1970s and 1980s, researchers constructed numerous instruments and questionnaires to assess organizational climate. Because of the belief that a healthy climate could be achieved and that it promoted numerous useful outcomes (Schneider, 1983; Victor & Cullen, 1988), researchers viewed climate as a promising tool for the analysis of organizational behavior. Representative of these instruments is the Charles F. Kettering Climate Profile (CFK), a popular measure that is widely used to gather educational data for organizational planning (Bailey & Young, 1989-1990; Dennis, 1979; Fox et al., 1973; Johnson, Dixon, & Johnson, 1992; Johnson, Dixon, & Robinson, 1987; Phi Delta Kappa, 1974). The instrument was patterned within the perceptual measurement-individual attribute approach to measuring and conceptualizing climate.

The purpose of this study was to clarify the appropriate conceptual variables and dimensions for the CFK instrument. This research will help in assessing perceived educational climate and in predicting individual behaviors and attitudes in educational settings. The suggested refinements for the Kettering scale are offered to help make the CFK more effective as a research instrument. Included in this article will be discussion of the dimensionality of climate perceptions and the psychometric properties of the Kettering scale. These topics all relate to a proper analysis of the Kettering scale.

Method

Subjects

A total of 1,311 administrators, teachers, staff, and students from an elementary, a junior high, and two high school campuses in a major school district in the southwestern United States completed Part A, the General Climate Factors section, of the profile. The subjects represented an available population that district personnel allowed to be assessed. The principals of the sampled schools were amenable to testing. For the 75 elementary students and personnel who completed the instrument, there were 21 fifth graders (age 10), 20 sixth graders (age 11), 6 teachers, and 28 support staff (2 secretaries, 4 adult aides, 10 university personnel, and 10 parents). Among the 257 junior high participants were 3 administrators, 12 teachers, 2 secretaries, 1 aide, 6 parents, 6 university personnel, 83 seventh graders (age 12), 66 eighth graders (age 13), and 78 ninth graders (age 17). In regard to the two high schools, there were 79 ninth and 79 tenth graders (ages 14 and 15) each from one school and 332 tenth graders (age 15), 249 eleventh graders (age 16), and 240 twelfth graders (age 17) from the other school. The range of the student ages was from 10 through 17 years. The mean age for the 1,247 students was 15 years with a standard deviation of 1.61. The sample was represented essentially equally by males and females. Females only slightly outnumbered males.

The elementary, junior high, and high school students' responses were pooled for this study since the instrument's developers developed their instrument to be administered and answered by teachers, administrators, and students collectively at all educational levels. Furthermore, analyses based on covariance require as much systematic variance as possible, to avoid range restrictions. The use of more diverse subject groups is potentially helpful in yielding the desired result. Out of the entire group of 1,311 subjects, there were only 64 adults. Therefore, the authors did not explore the congruence of the factors for students versus school personnel given the size of disproportionately small adult group.

Procedure

Subjects filled out the General Climate Factors profile in their school classrooms or work areas. They were not required to sign their names but were asked to complete a short demographic section indicating their status. The elementary and junior high students were from lower middle-class backgrounds, while the secondary students were from middle-class backgrounds. Most of the students were of Caucasian ancestry.

Instrument

The Kettering instrument was developed in the early 1970s (Johnson et al., 1992). Under the direction of R.S. Fox, a task force of educators associated with Charles F. Kettering II and his educational foundation researched the climate literature and developed the CFK. The CFK was copyrighted in 1973. The instrument was designed to be used in school settings. As a part of the CFK test development, the content validity was assessed by asking over 200 educators throughout the United States to respond to the instrument items (Johnson et al., 1992).

The CFK is composed of four sections: Part A, General Climate Factors (40 questions); Part B, Program Determinants (35 questions); Part C, Process Determinants (40 questions); and Part D, Material Determinants (15 questions) (Howard, Howell, & Brainard, 1987; Phi Delta Kappa, 1974).

The General Climate Factors section of the instrument consists of eight subscales. The number of items in each subscale follows: (a) Respect (Items 1-5), (b) Trust (Items 6-10), (c) High Morale (Items 11-15), (d) Opportunity for Input (Items 16-20), (e) Continuous Academic and Social Growth (Items 21-25), (f) Cohesiveness (Items 26-30), (g) School Renewal (Items 31-35), and (h) Caring (Items 36-40). See Figure 1 for a listing of the instrument questions.

The scaling technique used involves two discrepancy-format columns. The "What Is" column is the perceived actual status of the skill or attitude whereas the "What Should Be" column is the perceived desired status of the skill or attitude. The "What Is" column choices are placed on the left of the survey questions while the "What Should Be" choices are placed on the right. Each column has four descriptors: 1 = almost never, 2 = occasionally, 3 = frequently, and 4 = almost always.

Results

Data Analysis

We used the SAS principal components program (SAS Institute, Inc., 1986) to examine the construct validity of the General Climate Factors section of the CFK. A relevant question pertaining to performing a principal components analysis is if different factors will emerge if 1s are put in the main diagonal than if communalities are used. Gorsuch (1983) suggests that with 30 or more variables, the differences between solutions are likely to be small and lead to similar interpretations. Harman (1967) stated, "There is much evidence in the literature that for all but very small sets of variables, the resulting factorial solutions are little affected by the particular choice of communalities in the principal diagonal of the correlation matrix" (p. 83). Nunnally (1978) noted, "It is very safe to say that if there are as many as 20 variables in the analysis, as there are in nearly all exploratory factor analyses, then it does not matter what one puts in the diagonal spaces" (p.418). A somewhat conservative conclusion is that when the number of variables is moderately large, say larger than 30, and the analysis contains virtually no variables expected to have low communalities, that is 0.4, then practically any of the factor procedures will lead to the same interpretations (Stevens, 1986).

The claim for the so-called convergence of principal components and common factor analysis as the number of variables increases is correct, as long as the universe of variables to which the model is extended has a finite and fixed number of determinate common factors. The justification for performing a principal components analysis in this study was that there were a large number of variables having moderate communalities. The authors also performed two principal factor analyses to verify that the results were essentially equivalent. Therefore, we report the findings of our principal components analysis, noting that one would expect little difference between principal components with iterative communality analysis and principal factor analysis.

Because the CFK uses two discrepancy-format columns, we performed two separate first order principal components analyses (Stevens, 1986), one for the "What Is" left side of the scale and one for the "What Should Be" right side of the scale. Using the Kaiser (1960) criterion, the "What Is" analysis yielded six factors, while the "What Should Be" analysis isolated four factors. The prerotation eigenvalues for the "What Is" factors were 12.40, 2.37, 1.88, 1.31, 1.19, and 1.07. The prerotation eigenvalues for the "What Should Be" components were 17.67, 2.91, 1.57, and 1.13. Results of these solutions involve a first factor that might be characterized as a general or g factor. This is a factor with which most of the items were highly correlated and suggests the existence of a unidimensional factor structure. In general, the presence of a g factor does

Figure 1
Instrument Questions for the CFK Scale

Respect

1. In this school even low achieving students are respected.
2. Teachers treat students as persons.
3. Parents are considered by this school as important collaborators.
4. Teachers from one subject area or grade level respect those from other subject areas.
5. Teachers in this school are proud to be teachers.

Trust

6. Students feel that teachers are "on their side".
7. While we don't always agree, we can share our concerns with each other openly.
8. Our principal is a good spokesman before the superintendent and the board for our interests and needs.
9. Students can count on teachers to listen to their side of the story and be fair.
10. Teachers trust students to use good judgment.

High Morale

11. This school makes students enthusiastic about learning.
12. Teachers feel pride in this school and in its students.
13. Attendance is good; students stay away only for urgent and good reasons.
14. Parents, teachers, and students would rise to the defense of this school's program if it were challenged.
15. I like working in this school.

Opportunity for Input

16. I feel that my ideas are listened to and used in this school.
17. When important decisions are made about the programs in this school, I, personally, have heard about the plan beforehand and have been involved in some of the discussions.
18. Important decisions are made in this school by a governing council with representation from students, faculty, and administration.
19. While I obviously can't have a vote on every decision that is made in this school that affects me, I do feel that I can have some important input into that decision.
20. When all is said and done, I feel that I count in this school.

Continuous Academic and Social Growth

21. The teachers are "alive;" they are interested in life around them; they are doing interesting things outside of school.
22. Teachers in this school are "out in front," seeking better ways of teaching and learning.
23. Students feel that the school program is meaningful and relevant to their present and future needs.
24. The principal is growing and learning, too. He or she is seeking new ideas.
25. The school supports parent growth. Regular opportunities are provided for parents to be involved in learning activities and in examining new ideas.

Cohesiveness

26. Students would rather attend this school than transfer to another.
27. There is a "we" spirit in this school.
28. Administration and teachers collaborate toward making the school run effectively; there is little administrator-teacher tension.
29. Differences between individuals and groups (both among faculty and students) are considered to contribute to the richness of the school; not as divisive influences.
30. New students and faculty members are made to feel welcome and part of the group.

School Renewal

31. When a problem comes up, this school has procedures for working on it; problems are seen as normal challenges; not as "rocking the boat."
32. Teachers are encouraged to innovate in their classroom rather than to conform.
33. When a student comes along who has special problems, this school works out a plan that helps that student.
34. Students are encouraged to be creative rather than to conform.
35. Careful effort is made, when new programs are introduced, to adapt them to the particular needs of this community and this school.

Caring

36. There is someone in this school that I can always count on.
37. The principal really cares about students.
38. I think people in this school care about me as a person; are concerned about more than just how well I perform my role at school (as student, teacher, parent, etc.).
39. School is a nice place to be because I feel wanted and needed there.
40. Most people at this school are kind.

not mean that there is only one interpretable factor, but rather that there is a large overriding factor with additional factors reflecting nuances of the factor structure (Daniel, 1991).

One result of these analyses was a matrix of correlations among the factors. The interfactor correlation matrices can be factored just as the two 40 x 40 intervariable correlation matrices can be. This method is called second-order factor analysis. Kerlinger (1984) noted that "While ordinary factor analysis is probably well understood, second-order factor analysis, a vitally important part of the analysis, seems not to be widely known and understood" (p. xiv). It is important to realize that researchers often want to analyze data with second-order factor analysis, because various levels of analysis give different perspectives (Gorsuch, 1983; Johnson & Johnson, in press). As Thompson (1990, p. 579) explained, "The first-order analysis is a close-up view that focuses on the details of the valleys and peaks in mountains. The second-order analysis is like looking at the mountains at a greater distance, and yields a potentially different perspective on the mountains as constituents of a range. Both perspectives may be useful in facilitating understanding of data." Kerlinger (1984), Thompson and Borrello (1986), Thompson and Miller (1981), and Wasserman, Matula, and Thompson (1993) have presented examples of second-order factor solutions.

The decision to extract second-order factors was driven by the desire to conduct a higher-order analysis and by the finding that the first-order varimax solutions involved numerous multiple loadings, thus suggesting a first-order oblique solution as well as a second-order result. See Tables 1 and 2 for the first-order interfactor correlation matrices and the promax rotated factor pattern matrices.

Two second-order factors were extracted from both the "What Is" and "What Should Be" interfactor correlation matrices and rotated to the varimax criterion. See Table 3 for the second-order varimax rotated factor pattern matrices.

Second-order factors such as these, then, often are interpreted. However, Gorsuch (1983) argued that this is not desirable:

> Interpretations of the second-order factors would need to be based upon the interpretations of the variables. Whereas, it is hoped that the investigator knows the variables well enough to interpret them, the accuracy of interpretation will decrease with the first-order factors, will be less with the second-order factors, and still less with the third-order factors. To avoid basing interpretations upon interpretations of interpretations, the relationships of the original variables to each level of the higher-order factors are determined. (p.245)

Table 1
First Order Interfactor Correlation Matrices

	What Is Factors							What Should Be Factors			
	I	II	II	IV	V	VI		I	II	III	IV
I	-	47	30	43	46	09	I	-	65	50	39
II		-	22	40	40	08	II		-	63	45
III			-	10	36	11	III			-	56
IV				-	26	19	IV				-
V					-	04					
VI						-					

Note. Decimal points omitted.

Table 2
PROMAX Rotated Factor Pattern Matrices For "What Is" and "What Should Be" Scale Items

Item	**Scale**	**What Is Factor**						**What Should Be Factor**			
		1	2	3	4	5	6	1	2	3	4
1	Respect	0.152	-0.138	**0.459**	0.058	-0.043	0.270	-0.027	0.165	-0.070	**0.607**
2	Respect	0.014	0.063	**0.646**	0.085	-0.067	-0.122	0.031	-0.009	0.036	**0.644**
3	Respect	-0.174	0.034	**0.343**	**0.497**	-0.052	-0.056	-0.042	0.098	-0.049	**0.656**
4	Respect	0.016	-0.050	**0.599**	0.257	-0.119	-0.158	0.012	-0.051	-0.017	**0.744**
5	Respect	-0.026	**0.410**	**0.473**	-0.043	-0.125	0.028	0.013	-0.020	0.193	**0.592**
6	Trust	0.065	0.085	**0.512**	-0.099	0.085	0.251	0.104	-0.118	0.236	**0.482**
7	Trust	-0.017	0.067	**0.454**	0.045	0.132	0.121	0.009	0.126	0.110	**0.523**
8	Trust	-0.086	**0.686**	0.086	0.012	0.020	-0.074	-0.052	0.105	**0.572**	0.178
9	Trust	0.036	-0.021	**0.534**	0.029	0.232	0.106	0.048	-0.044	**0.606**	0.210
10	Trust	-0.162	**0.429**	**0.326**	0.082	-0.027	0.207	0.075	-0.136	**0.724**	0.003
11	High Morale	-0.004	**0.462**	0.125	-0.041	0.071	**0.412**	-0.026	0.066	**0.680**	0.116
12	High Morale	0.051	**0.636**	0.210	0.030	-0.143	0.165	0.041	0.019	**0.620**	0.115
13	High Morale	-0.029	0.096	0.049	-0.037	0.179	**0.720**	-0.032	0.000	**0.619**	0.156
14	High Morale	0.078	**0.650**	0.005	-0.010	-0.071	0.071	-0.032	0.205	**0.592**	0.034
15	High Morale	0.227	**0.462**	0.156	-0.296	0.238	-0.134	0.025	0.050	**0.641**	0.002
16	Input	0.061	-0.026	0.110	0.017	**0.658**	0.158	0.126	0.082	**0.633**	-0.087
17	Input	-0.056	-0.127	-0.038	0.078	**0.784**	0.235	-0.023	0.170	**0.619**	-0.101
18	Input	-0.094	**0.535**	-0.204	0.252	0.141	0.153	0.047	**0.404**	**0.474**	-0.092
19	Input	-0.011	0.094	-0.113	0.218	**0.658**	-0.057	0.103	**0.389**	**0.417**	-0.137
20	Input	0.259	0.273	0.005	0.009	**0.391**	-0.001	0.122	**0.343**	**0.405**	-0.007
21	Growth	0.021	-0.081	**0.321**	**0.375**	**0.365**	-0.134	0.049	**0.541**	0.196	0.042
22	Growth	0.078	0.145	0.263	**0.386**	0.190	-0.044	0.035	**0.640**	0.187	0.009
23	Growth	0.194	**0.340**	-0.144	0.295	-0.058	0.279	0.105	**0.658**	0.065	0.043
24	Growth	0.180	**0.396**	-0.026	0.255	0.145	-0.160	0.017	**0.762**	0.070	0.005
25	Growth	0.034	-0.025	0.003	**0.577**	0.257	-0.013	0.073	**0.691**	0.049	0.014
26	Cohesiveness	**0.461**	0.272	-0.140	0.096	0.029	-0.047	0.073	**0.658**	0.077	-0.133
27	Cohesiveness	**0.334**	**0.421**	-0.101	0.180	-0.092	-0.008	0.074	**0.676**	0.028	0.043
28	Cohesiveness	0.167	0.164	0.016	**0.533**	0.064	-0.038	0.048	**0.744**	-0.068	0.078
29	Cohesiveness	0.205	0.095	0.045	**0.473**	0.069	0.115	0.153	**0.660**	-0.045	0.095
30	Cohesiveness	**0.458**	0.192	0.036	0.258	-0.011	-0.077	0.255	**0.656**	-0.007	-0.003
31	Renewal	**0.345**	0.197	-0.086	**0.352**	0.036	-0.057	0.203	**0.629**	-0.001	0.001
32	Renewal	**0.485**	-0.204	0.116	**0.462**	-0.023	0.068	**0.595**	0.246	-0.027	0.022
33	Renewal	**0.643**	-0.062	0.059	0.170	-0.039	0.092	**0.708**	0.207	-0.040	0.034
34	Renewal	**0.617**	0.035	0.008	0.200	-0.145	0.138	**0.740**	0.102	0.003	-0.012
35	Renewal	**0.653**	0.121	-0.061	0.176	-0.131	0.041	**0.787**	0.100	0.032	- 0.041
36	Caring	**0.792**	-0.045	0.045	-0.026	0.008	-0.100	**0.821**	0.041	0.050	0.025
37	Caring	**0.718**	0.140	0.016	-0.057	0.019	-0.165	**0.799**	0.089	0.000	0.014
38	Caring	**0.815**	-0.105	0.034	-0.031	0.098	0.046	**0.874**	-0.022	0.040	0.006
39	Caring	**0.802**	-0.056	0.049	-0.157	0.160	-0.001	**0.875**	-0.054	0.055	-0.017
40	Caring	**0.816**	-0.038	0.021	-0.082	-0.001	0.078	**0.895**	-0.001	-0.006	0.028

Table 3
VARIMAX Rotated Second Order Factor Pattern Matrices for "What Is" and "What Should Be" Scale Items

What Is Factor		What Should Be Factor	
1	2	1	2
0.763	-0.216	-0.681	-0.648
0.662	-0.386	-0.830	0.087
0.167	0.876	-0.101	0.949
0.196	-0.835	0.978	0.066
0.725	0.328		
-0.950	-0.115		

The first-order promax rotated factors, therefore, were postmultiplied by the second-order varimax rotated factors, and the product matrices (for "What Is" and "What Should Be") were then rotated to the varimax criterion. Table 4 presents these factor pattern coefficients for items that had coefficients greater then 0.3. An approximate value for a statistically significant factor loading can be obtained by doubling the critical value required for an ordinary correlation. The statistically significant value for a sample size of 1000 is approximately 0.16 (Stevens, 1986). Very often in research, the value is set at 0.3 in absolute magnitude.

We used the generalized Kuder-Richardson reliability formula, coefficient alpha (Cronbach, 1951; Ebel, 1965; Novick & Lewis, 1967), to evaluate the reliability of the instrument. This formula was appropriate since a Likert scaling format was employed in the instrument form. The Cronbach alphas for the "What Is" factors (subscales) follow: subscale one (.91), subscale two (.71), subscale three (.82), and the composite for all "What Is" questions (.94). The Cronbach alphas for the "What Should Be" factors (subscales) follow: subscale one (.92), subscale two (.89), subscale three (.94), and the composite for all "What Should Be" questions, (.97). Although the CFK's developers never published reliability and validity date for their instrument, following are the Cronbach alpha values for the original subscales based on an analysis of the n=1311 data: "What Is" - Respect (.54), Trust (.63), High Morale (.66), Opportunity for Input (.74), Academic and Social Growth (.75), Cohesiveness (.80), School Renewal (.81), and Caring (.87); "What Should Be" - Respect (.73), Trust (.77), High Morale (.83), Opportunity for Input (.83), Academic and Social Growth (.88), Cohesiveness (.87), School Renewal (.88), and Caring (.93).

The subscale intercorrelations for the "What Is" subscales follow: (a) Factors one and two (.57); (b) Factors one and three (.82), and (c) Factors two and three (.62). The "What Should Be" subscale intercorrelations follow: (a) Factors one and two (.74), (b) Factors one and three (.78), and (c) Factors two and three (.63). These intercorrelations do not represent factor scores but subscale scores derived by summing the response category values for the salient items for a subscale.

Discussion

The findings presented in Table 2 indicate that the questions do not group as proposed by the instrument's developers. Table 2 data also show that there are six "What Is" first-order factors and four "What Should Be" first-order factors.

The factors presented in Table 4 indicate that two second-order factors represent the eight postulated scales for the Kettering instrument. The "What Is" column questions are comprised of 25 questions for factors one and two. The 10 items having factorial complexity are listed last in Table 4. Factor one is composed of 18 questions, and factor two is composed of 10 questions. Twelve questions were factorially complex in that these items correlated with both factors.

These findings suggest there are two second-order "What Is" subscales and two second-order "What Should Be" subscales. The first "What Is" subscale is a composite mainly of the last three sections of the CFK. The subscale is a composite of growth, cohesiveness, and school renewal questions. These questions are cognitive-managerial in nature. The second subscale is composed of questions from the respect, trust, opportunity for input, and growth sections of the instrument. The questions focus on affective-experiential components. The first "What Should Be" subscale focuses on cognitive-managerial features, while the second subscale focuses on affective-experiential components.

This analysis also suggests a student-related scale that measures student's feelings and perceptions of how they are treated and dealt with by teachers and by the school in general. This scale deals with climate from the perspective of students' lives in the school.

Table 4
Rotated Pattern Coefficients for Salient Items for "What Is" and "What Should Be" Scale Items

What Is				What Should Be			
		Factor				Factor	
Item	Scale	1	2	Item	Scale	1	2
1	Respect	**-0.344**	0.142	1	Respect	**-0.479**	-0.048
8	Trust	**0.484**	0.183	2	Respect	**-0.614**	-0.012
14	Morale	**0.467**	0.014	3	Respect	**-0.593**	-0.033
18	Input	**0.543**	-0.260	4	Respect	**-0.760**	-0.065
22	Growth	**0.412**	0.213	5	Respect	**-0.587**	0.147
24	Growth	**0.770**	0.167	6	Trust	**-0.491**	0.124
25	Growth	**0.497**	-0.096	7	Trust	**-0.402**	0.099
26	Cohesiveness	**0.707**	0.073	21	Growth	**0.434**	0.255
27	Cohesiveness	**0.700**	-0.074	22	Growth	**0.538**	0.272
28	Cohesiveness	**0.684**	-0.118	23	Growth	**0.572**	0.118
29	Cohesiveness	**0.471**	-0.164	24	Growth	**0.628**	0.194
30	Cohesiveness	**0.685**	0.108	25	Growth	**0.604**	0.120
31	Renewal	**0.726**	-0.050	26	Cohesiveness	**0.603**	0.151
32	Renewal	**0.410**	-0.082	27	Cohesiveness	**0.563**	0.104
33	Renewal	**0.436**	0.063	28	Cohesiveness	**0.565**	0.038
34	Renewal	**0.465**	-0.095	29	Cohesiveness	**0.559**	-0.015
35	Renewal	**0.638**	-0.038	30	Cohesiveness	**0.728**	-0.033
40	Caring	**0.446**	0.259	31	Renewal	**0.664**	-0.003
2	Respect	-0.103	**0.510**	8	Trust	-0.129	**0.586**
4	Respect	-0.044	**0.358**	9	Trust	-0.209	**0.534**
5	Respect	-0.003	**0.325**	10	Trust	-0.061	**0.623**
7	Trust	0.250	**0.411**	11	Morale	-0.082	**0.670**
9	Trust	-0.186	**0.503**	12	Morale	-0.069	**0.566**
16	Input	0.114	**0.454**	13	Morale	-0.179	**0.602**
21	Growth	0.288	**0.406**	14	Morale	0.107	**0.617**
6	Trust	**-0.310**	**0.370**	15	Morale	0.054	**0.606**
13	Morale	**-0.407**	**-0.311**	16	Input	0.241	**0.551**
15	Morale	**0.422**	**0.660**	17	Input	0.217	**0.638**
19	Input	**0.528**	**0.315**	18	Input	**0.382**	**0.555**
20	Input	**0.536**	**0.394**	19	Input	**0.525**	**0.415**
23	Growth	**0.455**	**-0.397**	20	Input	**0.375**	**0.379**
36	Caring	**0.573**	**0.370**	32	Renewal	**0.623**	**-0.319**
37	Caring	**0.697**	**0.415**	33	Renewal	**0.664**	**-0.403**
38	Caring	**0.481**	**0.327**	34	Renewal	**0.648**	**-0.398**
39	Caring	**0.469**	**0.482**	35	Renewal	**0.710**	**-0.396**
				36	Caring	**0.623**	**-0.411**
				37	Caring	**0.656**	**-0.437**
				38	Caring	**0.630**	**-0.462**
				39	Caring	**0.627**	**-0.452**
				40	Caring	**0.642**	**-0.514**

Note. Salient items were items with pattern coefficients greater in absolute value than .30.

Summary and Conclusion

This research analysis suggests that the currently used subscale subdivisions may be inappropriate. We understand from the CFK developers that they used only content validity in the instrument construction. The general test development literature suggests, however, that at least two types of validity measures should be used in scale development (American Psychological Association, 1985).

When the CFK developers departed from this traditional approach to instrument construction, they arbitrarily designated and assigned names to various subscales in their instrument. However, our first-order analysis did not verify the instrument developers' proposed structure. Our second-order solution found subscales that were cognitive-managerial and affective-experiential in nature. The suggested modifications for the Kettering scale are offered to help make the CFK more effective as a research instrument. Such is the nature of instrument refinement.

References

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Anderson, C. S. (1982). The search for school climate: A review of the research. *Review of Educational Research, 52*, 368-420.

Bailey, S. S., & Young, K. M. (1989-1990). The relationship between leadership styles of high school principals and school climate as perceived by teachers. *National FORUM of Education Administration and Supervision Journal, 6*, 108-123.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., II, & Weick, K. E., Jr. (1970). *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Daniel, L. G. (1991). Operationalization of a frame of reference for studying organizational culture in middle schools. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (pp.1-24). Greenwich, CT: JAI Press.

Dennis, J. D. (1979). *An assessment of the construct validity and the reliability of the CFK Ltd. School Climate Profile.* Unpublished doctoral dissertation, University of Colorado, Boulder.

Drexler, J. A. (1977). Organizational climate: Its homogeneity within organizations. *Journal of Applied Psychology, 62*, 38-42.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Forehand, G. A., & Gilmer, B. V. H. (1964). Environmental variation in studies of organizational behavior. *Psychological Bulletin, 62*, 361-382.

Fox, R. S., Boies, H. E., Brainard, E., Fletcher, E., Huge, J. S., Martin, C. L., Maynard, W., Monasmith, J., Olivero, J., Schmuck, R., Shaheen, T. A., & Stegeman, W. H. (1973). *School climate improvement: A challenge to the school administrator.* Bloomington, IN: Phi Delta Kappa.

Good, T. L., & Brophy, J. E. (1986). School effects. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Guion, R. M. (1973). A note on organizational climate. *Organizational Behavior and Human Performance, 9*, 120-125.

Halpin, A. W. (1966). *Theory and research in administration.* New York: Macmillan.

Harman, H. H. (1967). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press.

Howard, E., Howell, B., & Brainard, E. (1987). *Handbook for conducting school climate improvement projects.* Bloomington, IN: Phi Delta Kappa Educational Foundation. (ERIC Document Reproduction Service No. ED 290 211)

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin, 81*, 1096-1112.

Johnson, W. L., & Johnson, A. M. (in press). Using SAS/PC for higher order factoring. *Educational and Psychological Measurement.*

Johnson, W. L., Dixon, P. N., & Johnson, A. M. (1992). A psychometric analysis of the Charles F. Kettering climate instrument. *Psychological Reports, 71*, 1299-1308.

Johnson, W. L., Dixon, P. N., & Robinson, J. R. (1987). The Charles F. Kettering Ltd. school climate instrument: A psychometric anaylsis. *Journal of Experimental Education, 56*, 36-41.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kerlinger, F. N. (1984). *Liberalism and conservatism: The nature and structure of social attitudes.* Hillsdale, NJ: Erlbaum.

Miskel, C., & Ogawa, R. (1988). Work motivation, job satisfaction, and climate. In N. J. Boyan (Ed.), *Handbook of research on educational administration* (pp. 279-304). New York: Longman.

Moos, R. H. (1974). Systems for the assessment and classification of human environments: An overview. In R. H. Moos & P. M. Insel (Eds.), *Issues in social ecology.* Palo Alto, CA: National Press Books.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and reliability of composite measurements. *Psychometrika, 32,* 1-13.

Phi Delta Kappa. (1974). *School climate improvement: A challenge to the school administrator, an occasional paper.* Bloomington, IN: Phi Delta Kappa Educational Foundation. (ERIC Document Reproduction Service No. ED 102 665)

SAS Institute, Inc. (1986). *SAS user's guide: Statistics, statistical analysis system.* Cary, NC: Author.

Schneider, B. (1972). Organizational climate: Individual preference and organizational realities. *Journal of Applied Psychology, 56,* 211-217.

Schneider, B. (1973). The perception of organizational climate: The customer's view. *Journal of Applied Psychology, 57,* 248-256.

Schneider, B. (1983). Work climates: An interactionist perspective. In N. W. Feimer & E. S. Geller (Eds.), *Environmental psychology: Directions and perspectives.* New York: Praeger.

Schneider, B., & Bartlett, C. J. (1968). Individual differences and organizational climate I: The research plan and questionnaire development. *Personnel Psychology, 23,* 332-333.

Schneider, B., & Bartlett, C. J. (1970). Individual differences and organizational climate II: Measurement of organizational climate by the multi-trait-multi-rater matrix. *Personnel Psychology, 21,* 332-333.

Schneider B., & Hall, D. T. (1972). Toward specifying the concept of work climate: A study of Roman Catholic diocesan priests. *Journal of Applied Psychology, 56,* 447-455.

Stern, G. G. (1970). *People in context: Measuring person-environment in education and industry.* New York: Wiley.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Tagiuri, R. (1968). The concept of organizational climate. In R. Tagiuri & G. W. Litwin (Eds.), *Organizational climate: Explorations of a concept* (pp. 1-32). Boston: Harvard University, Division of Research, Graduate School of Business Administration.

Thompson, B. (1990). SECONDOR: A program that computes a second-order principal-components analysis and various interpretation aids. *Educational and Psychological Measurement, 50,* 575-580.

Thompson, B., & Borrello, G. M. (1986). Second-order factor structure of the MBTI: A construct validity assessment. *Measurement and Evaluation in Counseling and Development, 18,* 148-153.

Thompson, B., & Miller, A. H. (1981). The utility of "social attitudes" theory. *The Journal of Experimental Education, 49,* 157-160.

Victor, B., & Cullen, J. (1987). A theory and measure of ethical climate in organizations. In W. C. Frederick (Ed.), *Research in corporate social performance and policy* (pp. 51-71). Greenwich, CT: JAI Press.

Victor, B., & Cullen, J. (1988). The organizational bases of ethical work climates. *Administrative Science Quarterly, 33,* 101-125.

Wasserman, J. D., Matula, K., & Thompson, B. (1993, November). *The factor structure of the behavior rating scale of the Bayley Scales of Infant Development-II: Cross-sample, cross-sectional, and cross-method investigations of construct validity.* Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED Forthcoming)

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 47-51

Students and the First Amendment: Has the Judicial Process Come Full Circle?

Donald F. DeMoulin
Western Kentucky University

The rules of educational law that exist today are the result of a process of gradual evolution through formulation and interpretation of the legislative and judicial branches of government. However, those who must apply the law to particular cases must, of necessity, define the parameters of the interpretation. In the case of First Amendment rights, one interpretation of court rulings, namely Hazelwood, has posed a major area of concern and controversy regarding the earlier landmark interpretation of Tinker. This manuscript examines First Amendment rights and illustrates points of contention and conflict as educators try to cope with seemingly diametrically opposing interpretations.

First Amendment Rights

While there are certainly more practical areas of school law than that of students' rights of expression--principals are far more likely to be concerned, for example, with tort liability and labor relations matters--there is perhaps no area which has more Constitutional ramifications. It is in the First Amendment, after all, that the core notions of free expression are found:

> Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievance. (Lunenburg & Ornstein, 1991, p. 345)

The Supreme Court has held, furthermore, in *Gitlow v. People of the State of New York* (1925) that the amendment--as incorporated by the Fourteenth Amendment's due process protections--applies as to state "impairment" of those fundamental rights as well (Yudof, Kirp, & Levin, 1992).

Despite the seeming sweep of the First Amendment's language, however, it is clear that "no law" has never meant **no** law: Only two Supreme Court justices, Hugo Black and William Douglas, have ever taken that absolutist position. What has emerged as the interpretative theory of choice for the Supreme Court over the years has been the "preferred position balancing theory" (Pember, 1990, p. 46). That approach, which presumes that freedom of expression will prevail, nonetheless allows for competing rights to "win" in cases where those competing interests are significant enough. In *Near v. Minnesota* (1931), for instance, the Supreme Court ruled that the government may prohibit publication of some information during wartime, with national security in essence "winning" over the normally preferred right to freedom of the press (Yudof et al., 1992).

First Amendment Rights in the Schools

In the school setting, freedom of expression was not a real issue until the 1960s:

> For centuries, students were presumed to have few constitutional rights of any kind. They were regarded as junior or second-class people and were told it was better to be seen and not heard. Parents were given wide latitude in controlling the behavior of their offspring and when these young people moved into schools or other public institutions, the government had the right to exercise a kind of parental control over them; *in loco parentis*, in the place of a parent. (Pember, 1990, p. 73)

In 1969, however, the Supreme Court "first extended First Amendment protection to students . . . [and] at least 125 other court decisions followed that precedent, repeatedly overruling administrative censorship of student publications and other forms of campus expression" (Overbeck & Pullen, 1991). The landmark case, *Tinker v. Des Moines Independent Community School District* (1969) involved the wearing of black armbands by school

---

Donald F. DeMoulin is an Associate Professor in the Department of Educational Leadership at Western Kentucky University. Donald DeMoulin has published numerous articles and books in the field of education and conducts workshops nationwide on organizational development, performance fulfillment, effective leadership practices, and student safety and achievement. Please address correspondence to Donald F. DeMoulin, Department of Educational Leadership, 423A Tate Page Hall, Western Kentucky University, Bowling Green, KY 42101.

children in opposition to the war in Vietnam. There the court, in a 7-2 ruling, established the principle that students' First Amendment rights do not stop "at the schoolhouse gate" and that such students are not merely closed-circuit recipients of only that which the State chooses to communicate. Such symbolic expression, the court said, should especially be protected in the school setting:

> The classroom is peculiarly the "marketplace of ideas." The Nation's future depends upon leaders trained through wide exposure to that robust exchange of ideas which discovers truth "out of a multitude of tongues, [not] through any kind of authoritative selection." (Overbeck & Pullen, 1991, p. 431)

The court, concerned with public schools becoming "enclaves of totalitarianism," recognized that the students' rights to free expression were neither absolute nor co-extensive with those of adults. Such rights may be violated, it ruled, when exercising them would substantially interfere with officials' ability to keep discipline. Wearing armbands, however, did not do that.

It is hard, in retrospect, to over-emphasize the impact that the *Tinker* decision had on American jurisprudence during the ensuing years. While the Supreme Court's 1943 ruling in *West Virginia Board of Education v. Barnette* had established that public school students had First Amendment rights (in a case involving West Virginia's policy of expelling students who refused to salute the flag while reciting the Pledge of Allegiance), it was not until after *Tinker* that the proverbial legal floodgates opened. The best evidence of the case's vitality: the fact that it was judicially cited more than 2,000 times in school free speech cases in the 20 years following its release (Eveslage, 1990).

Apart from its quantitative impact upon the development of the law in this area, the case also had a qualitative influence on later court rulings all over the country; students actually began to win some of their litigation.

> But after *Tinker*, many more cases arose, as students asserted their newly won constitutional rights. Some of the earliest post-*Tinker* cases were only federal district court decisions and hence of limited value as precedents, but students were winning lawsuits against school officials. (Overbeck & Pullen, 1991, p. 435)

*The Impact of the **Tinker** Decision*

Courts following *Tinker*--until the mid-1980s, at least--tended to take the Supreme Court's language literally. High school administrators "could censor student speech only if it presented a genuine possibility of disruption, was libelous, was obscene, or promoted illegal activity" (Middleton & Chamberlain, 1991, p. 42). In Illinois, Indiana, and Wisconsin, in fact--since a 1972 ruling by the Court of Appeals for the Seventh Circuit in the case of *Fujishima v. Board of Education*--students "have had the same protection as the professional press against both prior review and censorship" (Eveslage, 1990, p. 23).

Despite the importance of *Tinker*, however, recent years have brought fresh problems and--particularly with personnel changes in the Supreme Court itself--new legal approaches. For some commentators, that is eminently logical:

> It is apparent now that the rather benign protest by a handful of students that resulted in the *Tinker* decision presented the courts with a rather simple problem to solve. Hence the seemingly broad constitutional protection erected in the 1969 ruling has not been fully carried forth when judges are faced with more complex issues. Courts have been increasingly reluctant to second-guess the actions of school administrators who are often viewed as being "on the firing line," forced to make quick decisions that may appear to be too stringent in hindsight. (Pember, 1990, pp. 73-74)

For others, though, recent judicial retrenchment from *Tinker* has wrought a situation in which "censorship and punishment for constitutionally protected student expression are common . . . [with] some school administrators . . . insensitive to constitutional values" (Gillmor, 1990, p. 635). That, in turn, has created an allegedly toothless high school press:

> Censorship is the fundamental cause of the triviality, innocuousness, and uniformity that characterize the high school press. It has created a high school press that in most places is no more than a house organ for the school administration. (Nelson, 1974, p. 4)

Significant Court Interpretations

There are two cases--both late-1980s rulings by the Supreme Court--which have drawn the most critical attention. One, *Bethel School District No. 403 v. Fraser* (1986) involved a speech by Tacoma, Washington, high school student Matthew Fraser nominating a friend for student body vice president. The speech, which was only six sentences long, contained sexual innuendoes but no profanity; while there was apparently no disruption apart from some student cheering and hooting, Fraser was nonetheless suspended from school for violating a school rule prohibiting the use of obscene language. He sued the school district and won in the Ninth Circuit Court of Appeals; he lost, however, in a 7-2 ruling, when the case was considered by the Supreme Court.

Chief Justice Warren Burger, in writing the majority opinion for the court, went to great lengths to emphasize that the court was not overturning its precedent in *Tinker*. "The undoubted freedom to advocate unpopular and controversial views in schools and classrooms," he noted, "must be balanced against the society's countervailing interest in teaching students the boundaries of socially appropriate behavior" (106 S. Ct. 3163, 1986, p. 49).

> Unlike the sanctions imposed on the students wearing armbands in *Tinker*, the penalties imposed in this case were unrelated to any political viewpoint. The First Amendment does not prevent school officials from determining that to permit a vulgar and lewd speech such as respondent's would undermine the school's basic educational mission. A high school assembly or classroom is no place for a sexually explicit monologue directed towards an unsuspecting audience of teenage students. (106 S. Ct. 3165, 1986, p. 51)

Justice John Paul Stevens, one of the two dissenters, emphasized the fact that Fraser's fellow students later selected him to be their commencement speaker as evidence that he had not violated their standards, that "he was probably in a better position to determine whether an audience composed of 600 of his contemporaries would be offended by . . . a sexual metaphor . . . than a group of judges who are at least two generations and 3,000 miles away from the scene of the crime (106 S. Ct. 3169, 1986).

As important as the *Fraser* decision was, it has not received anywhere near the critical scrutiny of its successor 1988 Supreme Court ruling, *Hazelwood School District v. Kuhlmeier* (Overbeck & Pullen, 1991). That case involved the censorship by a St. Louis-area principal of his high school's student newspaper; he removed two pages that dealt--using interviews with students whose names were not given--with teenage pregnancy (including discussions about abortion and birth control) and the impact of parents' divorces on their children. The principal argued, in defending his actions, that the stories would serve as invasions of privacy as to the individuals involved; further, he expressed concern that the stories were not editorially balanced.

The federal district court, in a suit by the students, ruled for the school; the appeals court, however, after finding that the *Tinker* standards had not been met--specifically, that there had been no showing that the censored articles would have either caused disruption or subjected the school to tort liability--found for the students. The Supreme Court reversed the appeals court by a 5-3 vote.

Rationale of Court Decision

The majority's decision in *Hazelwood* also did not purport to overturn the precedent established in *Tinker*; still, the justices made it clear that high school students are not adults for First Amendment purposes. Crucial to the majority's reasoning was the fact that the newspaper in question was published as part of a school journalism class. *Tinker* had dealt with armbands as personal expression that just happened to occur on school property; in *Hazelwood*, however, the newspaper was considered an official school-tool of communication--not an open forum--and that distinction was important to the five justices:

> Educators are entitled to exercise greater control over . . . [school-sponsored publications] to assure that participants learn whatever lessons the activity is designed to teach, that readers or listeners are not exposed to material that may be inappropriate for their level of maturity, and that the views of the individual speaker are not erroneously attributed to the school . . . we hold that educators do not offend the First Amendment by exercising editorial control over the style and content of student speech in school-sponsored expressive activities so long as their actions are reasonably related to legitimate pedagogical concerns. (484 U.S. 260, 108 S. Ct. 562)

Justice William Brennan, in a stinging dissent, criticized the school's "brutal censorship" and the principal's "unthinking contempt for individual rights." Outside critics of the decision have been no less vociferous in their conclusions about the effects of the ruling.

Conclusion

Cases like *Hazelwood* . . . are troublesome. If we are indeed educating our youth for citizenship, these holdings breed cynicism: free expression is not a right to be taken seriously. They assume that scholastic journalism has little role in making schools safer, healthier, and better places to be. Too often the school as an agent of government sends reverse constitutional messages to students when it represses dissent or unorthodox views. It is difficult to be optimistic about a ruling that will be broadly interpreted by school officials to condone censorship (Gillmor, 1990, pp. 645-646).

While the court's ruling could have been expected to find widespread support among school administrators, it has received support from some rather unexpected areas as well. The nation's regular daily press, for example, which one might expect to be critical of the decision, overwhelmingly agreed with the court's ruling. An *Editor & Publisher* survey found that papers who editorialized on the issue found the court's reasoning to be a logical extension of the principle that freedom of the press "is for those [here, the school and principal] who own one" (Pember, 1990, p. 75).

It is also important that the ruling be carefully analyzed for what it did not do. Schools still cannot routinely censor individual expression, nor--according to a Ninth Circuit Court of Appeals decision in 1988--would the decision apply to unofficial or "underground" newspapers. Also, the ruling does not affect the legal status of public university newspapers; it seems clear, in fact, that a public university "is constitutionally prohibited from controlling the content of its student publications" (Walden, 1988, p. 708).

It is also important to note that the court's decision in *Hazelwood* does not prohibit states from acting on their own to protect such student expression. California, for instance, in section 48907 of its Education Code, prohibits administrative censorship of school newspapers unless the content is obscene, libelous, or likely to cause disruption; Iowa and Massachusetts have since passed similar statutes (Overbeck & Pullen, 1991).

The long-term effects of *Hazelwood* are not, of course, yet apparent. Some have predicted that the decision "will surely encourage school principals to censor student newspapers, regardless of whether they are legally permitted to do so under the local rules" (Overbeck & Pullen, 1991, p. 445). Others, however, looking at the judicial consequences and noting that it took the lower courts a long time to adjust to *Tinker*, predict that those same courts may now move relatively slowly in expanding *Hazelwood* beyond its narrowest boundaries:

> Grudgingly or not . . . most courts by the 1980s were tempering a school's autonomy and acknowledging students' right to speak. Many of the lower courts, finally as comfortable with the prevailing judicial atmosphere of free expression as they were with school control in the 1960s, seem reluctant to change direction as abruptly as the Supreme Court's recent rulings suggest. Only tomorrow's history will show whether *Hazelwood*'s Supreme Court-sanctioned return to a more regressive public school atmosphere will prevail, or if lower court rulings will echo the 20-year litany of decisions that encouraged students to practice their citizenship. (Eveslage, 1990, p. 44)

Of all the freedoms guaranteed in this nation, none is more valued that the right of free speech and freedom of the press as set out in the First Amendment to the Constitution. However, these rights have yet to be ruled absolute.

As with the case with *Tinker*, some individuals thought that the schools would lose all authority in the classroom; this has not been the case. Only two years after *Tinker*, the Sixth Circuit, in *Guzick v. Drebus* (Overbeck & Pullen, 1991), upheld a school rule banning the wearing of freedom buttons distinguishing this case from *Tinker* in that, in this instance, the wearing of buttons and other insignia had a long-standing history of disruption and related to discipline in the schools (Hollander, 1978). However, with the ruling in *Hazelwood*, mixed messages have been sent to educators concerning individual interpretations.

As educators try to minimize their involvement in litigation, contradictory interpretations such as *Tinker* and *Hazelwood* (and a trio of seemingly dichotomous Supreme Court decisions related to student expression) have produced an atmosphere of uncertainty for administrators and teachers as they strive to maintain an ordered and disciplined educational climate. It may well be that, after 23 years, the judicial process has come full circle.

References

Bethel School District No. 403 v. Frazer, 106 S. Ct. 3163 (1986), p. 46.

Bethel School District No. 403 v. Frazer, 106 S. Ct. 3165 (1986), p. 51

Bethel School District No. 403 v. Frazer, 106 S. Ct. 3169 (1986).

Eveslage, T. (1990). First Amendment and high schools. *Media History Digest, 10*, 23.

Gillmor, D. M. (1990). *Mass communication law, cases, and comment* (5th ed.). St. Paul, MN: West.

Hazelwood School District v. Kuhlmeier, 484 U.S. 260, 108 S. Ct. 562.

Hollander, P. (1978). *Legal handbook for educators.* Boulder, CO: Westview Press.

Lunenburg, F. C., & Ornstein, A. C. (1991). *Educational administration: Concepts and practices.* Belmont, CA: Wadsworth.

Middleton, K. R., & Chamberlain, B. F. (1991). *The law of public communication.* New York: Longman.

Nelson, J. (1974). *Captive voices: The report of the Commission of Inquiry into High School Journalism.* Robert F. Kennedy Memorial.

Overbeck, W., & Pullen, R. (1991). *Major principles of media law.* Fort Worth, TX: Harcourt, Brace & Jovanovich.

Pember, D. R. (1990). *Mass media law* (5th ed.). Dubuque, IA: Brown.

Walden, R. (1988). The university's liability for libel and privacy invasion by student press. *Journalism Quarterly, 65*, 708.

Yudof, M. G., Kirp, D. L., & Levin, B. (1992). *Educational policy and the law.* St. Paul, MN: West.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 53-60

Metaphor Analysis: An Alternative Approach for Identifying Preservice Teachers' Orientations

Janet C. Richards
University of Southern Mississippi-Gulf Coast

Joan P. Gipe
University of New Orleans

Metaphor analysis may provide an alternative to more traditional survey methods for determining preservice teachers' beliefs about teaching. This qualitative inquiry examined the value of collecting and analyzing preservice teachers' pre- and post-semester metaphors about teaching by comparing the orientational content of the metaphors to the preservice teachers' teaching beliefs expressed in their journal entries and the language patterns they employed while teaching small groups of urban elementary students. The preservice teachers' metaphorical teaching orientations fell into two broad categories--Teacher as Information Giver and Student-Centered. These were consistent with their journal entries and language used while teaching. Metaphor and journal data and observations provide a rich source of information about preservice teachers' professional development since subtle changes in beliefs can be observed.

Traditionalists define metaphors as various types of widely-used figurative language which states an equivalence "between two separate semantic domains" (Sapir, 1977, p. 4) (e.g., "Our trip to Alaska was out of this world"). Current perspectives regard metaphors as representing our entire conceptual system including "the way we think, what we experience, and what we do everyday" (Lakoff & Johnson, 1980, p. 31). This more contemporary position also assumes that metaphors are generative. That is, how we perceive and metaphorically describe problems are central to how we address and generate solutions to those problems (Munby, 1986; Schon, 1979). For example, teachers may refer to students having difficulty in reading as "remedial students" or as "students in need of rich literacy experiences." Such descriptions have the dual capacity of revealing and influencing teachers' instructional practices. Thus, according to the more contemporary view, metaphors consciously and subconsciously define our realities, and may subtly guide our decisions (Lakoff & Johnson, 1980; Sibbett & Cawood, 1983).

Researchers interested in alternative ways to evaluate teachers' cognitions suggest that preservice teachers' metaphors may also indirectly reveal their previously acquired beliefs and conceptions about teaching.[1] Serving as a powerful filter, professional beliefs have the capacity to impact all aspects of preservice teachers' work in field placements, including what ideas and concepts they choose to accept or reject, and how they teach their lessons (Zeichner, Tabachnick, & Densmore, 1987). Unlocking the meaning of preservice teachers' metaphors may help to make their beliefs and conceptions "more explicit and accessible to analysis" (Bullough, 1991, p. 44).

Of course, a possibility exists that preservice teachers' metaphors represent nothing more than habitualized (i.e., "frozen") professional speech which is disconnected from their actual teaching views and practices (Aspin, 1984). All disciplines contain glib, metaphorical expressions which have lost their meaning over time, and the field of education is no exception (Pollio, 1987). However, some studies have documented "novel" and consistent metaphorical orientations in individual classroom teachers' descriptions of their work.

Janet C. Richards is Assistant Professor in the Department of Education and Psychology at the University of Southern Mississippi-Gulf Coast. Joan P. Gipe is Research Professor of Curriculum and Instruction at the University of New Orleans. Please address correspondence regarding the paper to Janet C. Richards, University of Southern Mississippi-Gulf Coast, 730 East Beach Blvd., Long Beach, MS 39560.

---

[1] Beliefs, orientations, and conceptions about teaching are defined as "the highly personalized ways in which a teacher understands classrooms, students, the nature of learning, the teacher's role in a classroom, and the goals of education" (Kagan, 1990, p. 423).

Categories include those labeled ontological, which refer to the mind, ideas, and the curriculum as objects (e.g., "His mind usually doesn't work"); a journey with a directionality (e.g., "Each child must start from the center and move up"); and a commodity passed through a conduit (e.g., "I have to get my message across") (Munby, 1987; Munby & Russell, 1989; Reddy, 1979). Other research suggests that teachers' "metaphors are related to [their] teaching practices" (Tobin, 1990, p. 122).

Thus, there is some support that metaphor analysis may provide a productive alternative to conventional methods traditionally employed to ascertain teaching beliefs (e.g., surveys, structured interviews, checklists). Yet, until recently, university teachers have largely ignored preservice teachers' figurative language, an exception being one study of students' metaphors in a year-long secondary teacher certification program (Bullough & Stokes, 1994). University teachers may be uncertain as to what exactly constitutes a metaphor, since there is no standard procedure for identifying and analyzing metaphors about teaching (Kagan, 1990). They may regard metaphors as "soft data" which cannot be measured, and therefore have no value (Eisner, 1988), or they may assume that preservice teachers' metaphors represent clichéd professional speech which demonstrates little semantic consistency or relationship to practice. The following qualitative inquiry: (a) explores the feasibility and value of collecting and analyzing preservice teachers' metaphors for the purpose of determining their current teaching orientations and (b) attempts to determine if preservice teachers' metaphorical statements relate to their teaching orientations as documented in their dialogue journals and in the language they employ while teaching.

Participants, Field Context, and Program Orientation

Participants were 23 female elementary education majors enrolled in a reading/language arts methods course block designed as an inquiry-oriented early field program and their two university teachers, who also served as observer/researchers. Studies suggest that when preservice teachers are confronted with teaching practices, values, and beliefs which "differ from their own . . . [they are] more likely to examine and reconstruct their own beliefs" (Kagan, 1992, p. 157). Therefore, all course activities (e.g., lectures, demonstration lessons, seminar discussions, preservice teachers working with small groups of students) were conducted two mornings a week in an urban elementary school specifically selected as the program context because of its nontraditional, permissive, student-centered atmosphere. For example, students in the school address teachers by their first names. They also feel free to socialize with peers during instructional sessions and are allowed to leave their classrooms in order to get a drink of water, walk in the hallways, or confer with the principal about their concerns and problems with teachers and friends.

The program was guided by a constructivist view of learning. For example, discussion topics included: (a) how human beings learn best (i.e., when they can explore, discover, reason, and continuously interact with their environment) and (b) the benefits of giving students some responsibility for their own learning (Harste, Short, & Burke, 1988; Vygotsky, 1986). The program also emphasized the importance of teachers reflecting about their work in order to attempt to solve educational problems in a thoughtful, deliberate manner (Dewey, 1933; Grossman, 1992). For instance, seminar topics focused on issues such as why all third graders must receive reading instruction from third grade basal readers or who ultimately is responsible if a student receives overly-harsh punishment from a teacher.

Research Methodology

Tenets of qualitative inquiry guided the research. Qualitative methods are especially appropriate when researchers wish to provide "rich, descriptive data about the contexts, activities, and beliefs of participants in educational settings" (Goetz & LeCompte, 1984, p. 17). According to Eisner (1991):

1. Qualitative studies tend to be field focused;
2. The researcher acts as an instrument;
3. The research is interpretive in nature;
4. The research makes use of expressive language and there is the presence of voice in the text. (pp. 32-41)

Data Collection: Part 1

During the first and last class meetings the university teachers reviewed the traditional definition of metaphors (i.e., inferential or figurative language connecting two dissimilar elements) and asked the preservice teachers to write a short metaphorical narrative describing their views about teaching and themselves as future teachers to include their current pedagogical beliefs and practices about how children learn best. The university teachers prompted the preservice teachers by saying, "Write your metaphor using your creativity. If you wish, you may begin your metaphor with one of these statements: (a) Teaching is like . . .; (b) Teaching and learning are like . . .; (c) Being a teacher is like . . . ."

Using the metaphor identification and recording system devised by Barlow, Kerlin, and Pollio (1971), the two university teachers independently analyzed and coded the preservice teachers' narratives looking for "novel" metaphorical content about teaching. Then, through discussion, the university teachers confirmed or settled differences in their opinions and interpretations of what constituted a metaphor until there was 100% agreement. Next, using constant comparisons (Glaser & Strauss, 1967; Tesch, 1990), the university teachers compared all of the metaphors and "listed, examined, and grouped them according to similar themes [or orientations]" (Weinstein, 1990, p. 281). Two distinct types of metaphors emerged in the narratives which the university teachers classified as "Teacher as Information Giver" and "Student-Centered." Two subcategories were further identified within the "Teacher as Information Giver" classification: "Curriculum as a Commodity" and "Curriculum as a Journey," with the latter apparently reflecting a less rigid orientation (See Appendix A for examples of the preservice teachers' metaphors and highlighted "signal" phrases).

Data Collection: Part II

The preservice teachers also wrote weekly journal entries and the two university teachers responded, particularly urging the preservice teachers to reflect about their teaching experiences (e.g., "You questioned why the students at this school address the teachers by their first name. Does this bother you? Why?"). Throughout the semester, the content of the journals was independently analyzed by the two university teachers, who again used a constant comparative method to identify statements which reflected teaching beliefs. The statements were coded for teaching orientations using the same categories which emerged in the preservice teachers' initial metaphors. For example, "Today I told them they were going to make pudding and then they were going to dictate a language experience story about it. I finally had to tell them what to dictate because they got all mixed up. Then I told them some of the sight words in the story that they needed to know," was coded as "Teacher as Information Giver." Once again, differences of opinions between the university teachers were resolved by discussion until 100% agreement was reached (See Appendix B for examples of the preservice teachers' journal entries and highlighted "signal" phrases).

Data Collection: Part III

Over the course of the semester the two university teachers independently observed the preservice teachers as they taught their lessons. Each preservice teacher was observed on at least eight occasions. Because both oral and written language reveal teachers' modes of thinking (Clift, Houston, & Pugach, 1991; Munby, 1986), the university teachers used a researcher-devised coding system to document the language the preservice teachers expressed as they taught small groups of students. Guided by the same categories which emerged in the preservice teachers' initial metaphors and journal entries, the university teachers coded the preservice teachers' language. For example, "Encourages students to voice their opinions" was coded as teacher language indicative of student-centered beliefs. As before, differences of opinion were settled until 100% agreement was reached (See Appendix C for an example of this coding system).

Data Analysis

At the end of the semester the university teachers collated the three data sets for each preservice teacher (i.e., pre- and post-semester metaphors, journal entries, and instructional language documented on the teacher observation checklist). The aggregated data for each preservice teacher were scanned, compared, and cross-checked in order to identify "categories of phenomena and . . . relationships among categories" (LeCompte & Preissle, 1993, p. 254). The university teachers looked for content commonalities and orientational consistency, or what Guba calls "recurring regularities" (1978, p. 204) in the preservice teachers' protocols. Thus, three different sources of information provided a tri-dimensional perspective of the preservice teachers' beliefs about teaching (Morine-Dershimer, 1983; Tesch, 1990). In addition, data from the journals and the observation coding system served to check the orientational consistency of the preservice teachers' metaphors.

Results

Each of the preservice teachers' pre- and post-semester narratives contained "novel" metaphors about teaching which demonstrated an orientational consistency throughout each narrative. The pre- and post-semester metaphors about teaching fell into two categories which the researchers labeled (a) "Teacher as Information Giver" (e.g., "Students who learn the most pay close attention to the teacher") and (b) "Student-Centered" (e.g., "You learn from your students like you learn different customs traveling through Europe"). However, two distinct subcategories emerged within the "Teacher as Information Giver" orientation. The researchers titled these subcategories (a) "Curriculum as a Commodity" (e.g., "I'll give it to them and make them want to buy it") and (b) "Curriculum as a Journey" (e.g., "I will lead the children on their trip through the forest of knowledge"), with the former apparently reflecting a more rigid view.

Representative examples of the preservice teachers' narratives are included below. The researchers' classifications of the metaphors and identified "signal phrases" follow each narrative.

The Restaurant. Restaurant clients (i.e., students) come in to eat junk foods just to get by. The teacher, as the cook, knows their nutritional needs and supplies them with the proper foods so that they will be healthy and able to function properly. You can see how well the customers (i.e., students) are doing by observing the increase in their health and how much better they are able to function because of the nutritious food that you serve them. (**Metaphor Classification**: "Teacher as Information Giver/Curriculum as a Commodity") (**Signal Phrases**: (a) "teacher . . . knows their nutritional needs and supplies them"; (b) . . . how much better they are able to function because of the nutritional food that you serve them")

A Guided Tour. I see the teaching and learning of children as a guided tour of a far away place. They have never experienced the wonderful things they will encounter and they need a leader to point things out and show them the way. The job of the tour guide is to lead the tourists to learn new things. (**Metaphor Classification**: "Teacher as Information Giver/Curriculum as a Journey") (**Signal Phrases**: (a) "teaching and learning . . . as a guided tour of a far away place"; (b) "The job of the tour guide is to lead the tourists to learn new things.")

The Human Body. I think that teaching is like the human body. My function would be that of the head or brain, taking in all the information and feelings of the rest of the body (students). Without my arms and legs (students) telling me what they need or want to do, I would have no direction or idea of how to assist. Without my body parts my brain would be stagnant. The brain and parts need each other for survival. (**Metaphor Classification**: "Student-Centered") (**Signal Phrases**: (a) Without my . . . students telling me what they need or want to do I would have no direction or idea of how to assist"; (b) " . . . need each other for survival").

Initially, 19 preservice teachers held "Teacher as Information Giver" views (9 preservice teachers "Curriculum as a Commodity"; 10 preservice teachers "Curriculum as a Journey"), and 4 preservice teachers held "Student-Centered" views. By the end of the semester 4 preservice teachers continued to hold a "Curriculum as a Commodity" view; 3 moved to a "Curriculum as a Journey" view, and 2 moved to a "Student-Centered" view. Of the 10 preservice teachers initially holding a "Curriculum as a Journey" view, none adopted the more rigid "Curriculum as a Commodity" view; 7 continued to hold a "Curriculum as a Journey" view, and 3 moved to a "Student-Centered" view. All of the 4 preservice teachers initially holding "Student-Centered" views continued to hold those views.

Additionally, there was consistency among the preservice teachers' teaching orientations as indicated in their metaphors, language expressed in journal entries, and language employed while teaching. That is, preservice teachers whose orientations remained stable throughout the semester wrote pre- and post-semester metaphors which were consistent in orientation. They also wrote journal entries and used instructional language which reflected those views. On the other hand, five preservice teachers wrote "Teacher as Information Giver" pre-semester metaphors and "Student-Centered" post-semester metaphors. By mid-semester, subtle changes demonstrating a "Student-Centered" view were noted in their journal entries and in their instructional language.

Discussion

The study reported here explores an alternative means for determining preservice teachers' teaching orientations. Caution must be used in drawing conclusions from this study, since teaching beliefs are the result of a complex set of variables including school context conditions. The possibility exists that preservice teachers "might employ different metaphorical figures at different times and under different circumstances" (Munby, 1986, p. 201). Therefore, generalizing the study's findings to other preservice teachers working in different school contexts with different university teachers is limited. Nonetheless, the results of the study present sufficient evidence that metaphor analysis is both a feasible and valuable means of documenting preservice teachers' orientations as well as their professional development.

The study shows that metaphor analysis can provide university teachers with an innovative and practical approach for identifying preservice teachers' teaching orientations. If solicited early in preservice students' teacher education programs, metaphor analysis affords an opportunity for university teachers to plan appropriate course activities and seminar discussions for nurturing preservice teachers' growth toward views more conducive to student learning. That is, if a preservice teacher is identified through his/her metaphor as holding a rigid transmission of knowledge orientation, the university teacher can present specific scenarios, pose teaching dilemmas, frame questions, and provide experiences which encourage preservice teachers to become aware of their beliefs and to consider reconceptualizing their teaching roles. Additionally, metaphor data coupled with data from journals and observations can provide university teachers with an even richer source of information about preservice teachers' professional development, since subtle changes in beliefs are visible in preservice teachers' language used in these ongoing activities. End of semester metaphor analysis can then be used to confirm these changes.

In this study, the preservice teachers were not informed about their teaching orientations as documented by their metaphors. However, the next step in this line of research would be to "apply what we have learned about metaphors [and preservice teachers'] beliefs" (Tobin, 1990, p. 126). Through the use of metaphor analysis, university and preservice teachers can work together and take an active role in examining their teaching beliefs.

References

Aspin, D. (1984). Metaphor and meaning in educational discourse. In W. Taylor (Ed.), *Metaphors of education* (pp. 21-37). London: Heinemann Educational Books for the Institute of Education, University of London.

Barlow, J., Kerlin, J., & Pollio, H. (1971). *Training manual for identifying figurative language.* Knoxville: University of Tennessee, Dept. of Psychology.

Bullough, R. (1991). Exploring personal teaching metaphors in preservice teacher education. *Journal of Teacher Education, 42*(1), 43-45.

Bullough, R., & Stokes, D. (1994). Analyzing personal teaching metaphors in preservice teacher education as a means for encouraging professional development. *American Educational Research Journal, 31*(1), 197-224.

Clift, R., Houston, W., & Pugach, M. (1991). *Encouraging reflective practice in education: An analysis of issues and programs.* New York: Teachers College, Columbia University.

Dewey, J. (1933). *How we think.* Lexington, MA: D. C. Heath. (Original work published in 1909.)

Eisner, F. (1988). Can educational research inform educational practice? *Phi Delta Kappa, 65*, 447-452.

Eisner, F. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practices.* New York: Macmillan.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine.

Goetz, J., & LeCompte, M. (1984). *Ethnography and qualitative design in educational research.* New York: Academic Press.

Grossman, P. (1992). Why models matter: An alternative view on professional growth in teaching. *Review of Educational Research, 62*(2), 171-179.

Guba, E. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation.* Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles Graduate School of Education.

Harste, J., Short, K., & Burke, C. (1988). *Creating classrooms for authors.* Portsmouth, NH: Heinemann.

Kagan, D. (1990). Ways of evaluating teacher cognition: Inference concerning the Goldilocks principle. *Review of Educational Research, 60*(3), 419-469.

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research, 62*, 129-169.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

LeCompte, M., & Preissle, J., with R. Tesch. (1993). *Ethnography and qualitative design in educational research.* New York: Academic Press.

Lortie, S. (1975). *Schoolteacher.* Chicago: University of Chicago Press.

Morine-Dershimer, G. (1983). *Tapping teacher thinking through triangulation of data sets.* Austin: Communication Services, Research and Development Center for Teacher Education, The University of Texas.

Munby, H. (1986). Metaphor in the thinking of teachers; An exploratory study. *Journal of Curriculum Studies, 18*, 197-209.

Munby, H. (1987). Metaphor and teachers' knowledge. *Research in the Teaching of English, 21*, 377-397.

Munby, H., & Russell, T. (1989). *Metaphor in the study of teachers' professional knowledge.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 309 164)

Pollio, H. (1987, Fall). *Teaching-learning issues: Practical poetry: Metaphoric thinking in science, art, literature, and nearly everywhere else.* Knoxville: The Learning Research Center/The University of Tennessee.

Reddy, M. (1979). The conduit metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284-324). Cambridge, England: Cambridge University Press.

Sapir, J. (1977). Anatomy of a metaphor. In J. Sapir & J. Crocker (Eds.), *The social use of metaphor* (pp. 3-32). Philadelphia: University of Pennsylvania.

Schon, D. (1979). Generative metaphor: A perspective on problem setting in social policy. In A. Ortony

(Ed.), *Metaphor and thought* (pp. 254-283). Cambridge, England: Cambridge University Press.

Sibbett, D., & Cawood, D. (1983). Multi-image metaphors for the management of meaning. In E. Kosower (Ed.), *Beyond our boundaries: Presenters' papers* (pp. 154-158). Los Angeles: Organization Development Network Conference.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. New York: The Falmer Press.

Tobin, K. (1990). Changing metaphors and beliefs: A master switch for teaching? *Theory into Practice, 29*(2), 122-127.

Vygotsky, L. (1986). In R. Reiber & A. Carton (Eds.), *The collected works of L. S. Vygotsky: Vol. 1. Problems of general psychology* (N. Minick, Trans.). New York: Plenum.

Weinstein, C. (1990). Prospective elementary teachers' beliefs about teaching: Implications for teacher education. *Teaching and Teacher Education, 6*, 279-290.

Zeichner, K., Tabachnick, R., & Densmore, K. (1987). Individual, institutional, and cultural influences on the development of teachers' craft knowledge. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 21-59). London: Cassell.

Appendix A

Examples of the Preservice Teachers' Metaphors and Highlighted "Signal" Phrases

Examples of Metaphors Coded as "Teacher as Information Giver/Curriculum as a Commodity"

- "Being a teacher is like being a **salesman**. You have to give a good **sales pitch** and get them interested enough to **buy the product**."

- "Teaching is like a trip to the circus. You are the **ring leader** and you **introduce** many interesting things to the students that they **hopefully will never forget**. The students have to **pay close attention** in order for them to be able to **absorb everything** that is going on around them. As the **ring leader**, I will **introduce** the acts and **explain** how the acts will be accomplished. The students will then leave and **be able to perform the acts they have learned**."

- "This may sound strange but teaching is like shopping for groceries and cooking. When you are shopping you usually start at the first aisle and go down putting **products** in your basket. Like shopping and cooking a meal, in teaching **you have to have all of the ingredients** to make it complete. You as a teacher have to **make sure you give your students all the right products to make them learn**."

- "When I think of teaching I think of a large body of water. The water represents a large **body of knowledge**. I will **help the children learn this knowledge** and **decide if they have learned it before they can go on** to the next ripple or level of knowledge."

- "The teacher teaches children **basic skills** like lifeguards teach children how to swim. The lifeguard gives swimmers a **skills test** about four times a year to find out **which level they are** in swimming. Teachers also **test to see if children have learned what they have been taught**."

Examples of Metaphors Coded as "Teacher as Information Giver/Curriculum as a Journey"

- "Teaching and learning are like climbing a mountain. Sometimes it's hard to reach the top but the **teacher who is the guide** is always there to help the kids through the **rough spots**. At the end the students **reach the top with a lot of help from the teacher**."

- "Teaching is like trying to **give directions** on finding a street to foreigners who don't speak English. You have to **draw them a map and label the key locations** and **show them step-by-step where to go**. If they choose to use the map they will be successful. If not, . . . they'll have to get **directions over and over again**."

- "Teaching is like a tour guide **leading a group** of tourists through a far away land. Everything is **mapped out** for them. The job of the tour guide is to **lead the tourists to learn new things** and **guide them in this learning process**."

- "I think of **learning as a walk down a path** in the woods. The path is winding and you are **led by a tour guide**. **The trail leader points out** the intricacies of the forest and its species. The principal has the role of caretaker. He must **keep the paths clear so the way will be smooth**."

- "Teaching is like a **path that leads some place** really great. The only problem is that the kids don't know how nice the place is and they may not even care. The **path is cluttered** and is often **difficult to make the way**

through. Some kids are having so many problems overcoming the **obstacles in their paths that it is impossible for them to move forward**. The teacher must stop and **help that child along the path** so that he is not left behind."

Examples of Metaphors Coded as "Student-Centered"

- "Teaching is really like **sharing** between the teacher and the kids and vice versa."

- "I'll be like an anthropologist. Anthropologists live among, and **almost become, those who they study**. They look at people from the society the people are from . . . not from the anthropologist's society. I will be an anthropologist and **look at things from my students' views**."

- "Teaching is like a balancing scale. In order for the scale to work or balance **there must be give and take**. Applied to teaching this means that teachers must look at students and **get their input**. Thus, it's a **balance of learning** . . . **teacher and students together**."

- "Teaching is always changing as the teacher **learns about her students' needs**. Just like a sculptor who discovers what's within a piece of clay, teachers discover that **each student is unique and different**. Then, the teacher and students **become "one" together**. Just like a sculptor with clay, it is **"we" instead of "me**."

- "Teaching is like a stepper, a piece of exercise equipment. It doesn't work if you don't use both pedals. **You aren't teaching if your students aren't learning** and **you only learn about your students if you know that they can teach you**. So, the two pedals are like a teacher and her students. **They need each other or it won't work**."

Appendix B

Examples of the Preservice Teachers' Journal Statements and Highlighted "Signal" Phrases

Examples of Journal Statements Coded as "Teacher as Information Giver"

- **"I had them tell me** how to prepare the brownie mix. **Then I told the students** to dictate a story orally. **I had each student give me** at least two sentences. **I had to give a boy a warning because he wanted to give me more than two sentences. I told them to listen to me** as I read the story."

- "During the lesson today **I had one student who gave me problems**. I hope he **won't bother me in the future**. He must learn to **listen to what I say**."

- **"I had them** construct a collage. **Then I told them** to make up a story to go along with it. **Maybe I should have let them speak when they wanted to. But, no one would learn anything then**."

- "It became clear that **these second graders are unable to write stories. They should be able to** write some type of story. **They don't know how to organize their thoughts. This is terrible**."

- "I was disappointed because **they did not conform to the rules I set for them. I hope those five behavioral problem students will realize** that good behavior is better than bad."

Examples of Journal Statements Coded as "Student-Centered"

- "I love my sixth graders. Today we made no-cook pudding. I know **we all** need practice on this. **We all** got a laugh when I dropped the spoon in the pudding. What fun."

- **"Thanks to Annette, one of my students, I found out a lot. She taught me** that you can't judge a person by their neighborhood."

- "If you're going to work with kids then **you must accept the nature of children**. You can't expect them to sit up and listen like adults in medical school."

- "I thought **they did a terrific job**. Randy didn't show up so I filled in for her. **I hope the kids don't mind**. Before going on the stage **we all took three deep breaths together**. Anyway, I thought they were just terrific."

- "Today was such a great day! The eighth graders and I had such a good time. Of course, **when I say the eighth graders and I, I mean we, us together**."

- "At some point in the day **they need time to do what they choose. They need choices of things to do and**

stuff that has real meaning to them. Teachers need to understand that."

Appendix C

Example of the Coding System Used to Document the Preservice Teachers' Instructional Language

Name ______________________ Date __________

"Teacher as Information Giver" Instructional Language

1. Strongly emphasizes procedures (e.g., "I am going to pass out the papers. When you receive your paper begin working. Your job is to answer all of the questions. When you have finished, turn in your paper.")
2. Continually tells students to be quiet (e.g., "Listen carefully."; "Pay attention to me."; "Don't socialize."; "Be quiet."; "Listen to directions."; "Listen to what I say."; "Don't share answers."; "You learn by listening.")
3. Interrupts and cuts off students' verbalizations (e.g., "That's enough."; "Okay, no more talking.")
4. Uses "I" statements rather than "we" statements (e.g., "I want you to finish your work."; "Today I'm going to teach you how to complete a cloze passage.")
5. Corrects students' academic responses in a harsh or terse manner (e.g., "No, you're wrong again."; "No"; "Think before you blurt out an answer.")
6. Uses a large amount of "teacher-talk" as opposed to encouraging student discussion

"Student-Centered" Instructional Language

1. Praises students often for trying (e.g., "You did a great job."; "Good thinking."; "Great idea.")
2. Encourages students to voice their opinions (e.g., "What do you think?"; "What ideas do you have?")
3. Acknowledges s/he is a learner along with students (e.g., "I never knew that."; "You really taught me something."; "I never thought of that.")
4. Encourages students to take some responsibility for their own learning and for organizing class activities (e.g., "Let's decide how we can accomplish this. Who has some ideas?")
5. Encourages student discussion and collaboration (e.g., "Get together and see what ideas you can come up with.")
6. Uses "we" statements rather than "I" statements (e.g., "We all need to understand how to do this.")

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 61-75

The Effects of Violations of Data Set Assumptions When Using the Oneway, Fixed-Effects Analysis of Variance And the One Concomitant Analysis of Covariance

Colleen Cook Johnson and Ernest A. Rakow
University of Memphis

This study integrates into one Monte Carlo simulation an array of studies into the robustness of the analysis of variance (ANOVA) and covariance (ANCOVA). Three sets of balanced designs and one set of unbalanced designs were simulated. Data set violations include skew and kurtosis, heterogeneity of variances, and with ANCOVA, heterogeneity of slopes and a skewed covariate. Each violation was simulated both in isolation and in combination with others, resulting in 665 empirical F distributions which were then compared to the nominal F distribution. The unbalanced designs produced statistically invalid F ratios with almost any violation. In the balanced design, however, robustness greater than suggested by Glass, Peckham, and Sanders (1972) was found. This finding is important because the most common violation in balanced designs is heterogeneity of variances. If a researcher finds the dependent's skew and kurtosis fall within the 95% confidence bands, then the variance ratio can be as high as 5 without jeopardizing the results.

Within a scientific discipline, theories unify the existing knowledge base as well as provide hypotheses for further extension of that knowledge base. Theories are abstractions, and, as such, are represented by the construction of conceptual or mathematical models, models which serve to abstract the subject under study while preserving the original structure of the system. By abstracting the subject into a succinct, parsimonious model, it is possible to determine how changes in one (or more) parts of a model might affect the system as a whole. Oftentimes these changes are impossible to observe and document in the real world; yet by manipulation of the model it is possible to shed light on both the effects of such change and the functioning of the model itself.

There are two types of models that can be developed: deterministic and probabilistic. Deterministic models are defined so that virtually 100% of the variance in the dependent variable can be explained by the independent variable(s) included in the model. For instance, "$E=mc^2$" can be considered a deterministic model if it can produce accurate estimates of E with little or no error. These models are seldom used in education, psychology, or the social sciences. Concerning this, Lord and Novick (1968) wrote, "Deterministic models have found only limited use in psychology . . . because for problems of any real interest . . . we are unable to write an equation such that the residual variation in the dependent variable is small" (p. 23).

Instead, probabilistic models are more common in these disciplines. These models are not powerful enough to eliminate unexplained variation, although strategic methods are often used to minimize the proportion of unexplained variance while maximizing the amount explained. The general linear model (GLM) is a classic example of a probabilistic model. It has been argued that use of one specific form of the GLM, the analysis of variance, is the most widely used statistical procedure in several major educational journals, and is widely used in the psychological and social science literature as well (Elmore & Woehlke, 1988; Goodwin & Goodwin, 1988; Halpin & Halpin, 1988). Like other statistical models, those who use the GLM must assume that the prerequisite conditions for using the model actually do exist within their data set. However, a researcher seldom stumbles into a situation where all prerequisite conditions are perfectly met. Therefore, it is necessary to examine the statistical model itself, in its various forms, to determine to what extent real world conditions may depart from the assumptions inherent in the model before the GLM should be abandoned in favor of other statistical models.

The Nature of Monte Carlo Experimentation

There are two different kinds of mathematical research: theoretical and experimental. The main concern of theoretical mathematics is abstraction and generality. The theoretical mathematician will write arguments in the form of symbolic expressions or formal equations which will abstract the essence of a problem, thus revealing the underlying structure. However, this strength is also its inherent weakness: The more general and formal the language, the less able the theory is at providing a numerical solution to a specific situation (Hammersley & Handscomb, 1967). The Monte Carlo approach allows the exploitation of the strengths of theoretical mathematics while avoiding the weaknesses inherent in it. Using this approach, experimentation replaces theoretical exploration when the latter falters.

Using Monte Carlo Simulation to Explore the Robustness of the General Linear Model

The GLM possesses a number of different forms, all of which provide an abstracted and succinct statement of the relationship between variables carefully chosen by research practitioners to reflect real world phenomena (Cohen, 1968; Knapp, 1978). Though the model is frequently used, the data collected for analysis never perfectly adhere to all of the assumptions of the model. Thus it becomes a question of how much difference there is between the conditions the model was designed to handle and the actual conditions that exist in a particular research situation. If the difference is within a "tolerable range," then use of one of the forms of the GLM should produce information that is statistically robust in its treatment of the relationship between variables. It is only when the data collected exceed that "tolerable range" that alternatives to the GLM must be considered. Theoretical mathematics can be used to define the general nature of the problems that emerge when the GLM is used inappropriately; however, it is unable to provide us with the precise limits of this "tolerable range."

Monte Carlo simulation provides valuable supplementary information about the problems that develop when assumptions underlying the GLM are violated. Using this methodology, it is possible to numerically define the degree of tolerance (i.e., robustness) that specific forms of the GLM have under real world research conditions.

This research is an empirical study of the effects of violations of the assumptions for two specific forms of the general linear model: the oneway, fixed-effects analysis of variance and the analysis of covariance using one independent variable and one concomitant (i.e., covariate). These two methods are used extensively in educational and psychological research, and serve as the mathematical foundation for more complex extensions of the GLM as well.

Unique Contributions of this Research

This study offers three unique contributions to the existing literature studying the appropriate use of ANOVA and ANCOVA in educational and psychological research. First, this study directly tested Harwell, Hayes, Olds and Rubinstein's claim (1990, 1992) that inflated Type I error rates result when the ratio of largest to smallest group variances in the balanced design is as small as two against the standard established by Glass, Peckham and Sanders (1972) that in balanced designs one need only be concerned about the effects of heterogeneity of variances if the ratio of largest to smallest variances is at least three. Second, the study combined a number of different violations both separately and in combination, thereby examining the effects of data set violations at the zero, first, second, third, and fourth orders. Most previous studies have been limited to exploration at the zero and first orders only. Finally, this study allowed for the systematic control of random noise that has confounded the results of past studies--thus providing findings that are more precise than those found in previous simulations.

Review of the Literature

The simplest prototype of the general linear model (GLM) is the t test for two independent samples, which tests for mean differences between two groups. The oneway, fixed-effects analysis of variance (ANOVA) is the logical extension of this t test, broadened in form to allow for the analysis of two or more groups. Both of these statistical procedures involve analysis of the effects of one discrete independent variable on a single, continuous dependent. In the oneway ANOVA, F represents the ratio of the variance in the dependent variable that can be explained by the researcher's data to that variance left unexplained. The analysis of covariance (ANCOVA) is a logical extension of the oneway ANOVA, applicable when a third, continuous variable (referred to as the covariate or concomitant variable) is known to have a significant effect on the dependent variable, while having little or no effect on the independent variable. When ANCOVA is appropriate, the researcher's goal is to probe the effects of the independent variable on the dependent after removing the influence of the concomitant. To do this, ANCOVA first removes all variation in the dependent variable that is a function of the concomitant. Then, using these "adjusted scores," ANCOVA effectively reanalyzes the data for mean differences between the groups that make up the independent variable.

The two forms of the GLM studied in this simulation are the oneway, fixed-effects ANOVA and the one concomitant ANCOVA. Most researchers in this area accept the premise proposed by Cochran (1957) and Winer (1962), who have claimed that the relationships found in the simple oneway ANOVA and even the more basic t test for two independent samples carry over into the ANCOVA extension. Therefore, this literature review will contain discussion of relevant theoretical and empirical research involving the use of all of these statistical models.

Statistical Models and the Assumptions Inherent Within Them

When they are initially developed, statistical models (i.e., procedures) are designed to be used under a specified set of conditions (that is, assumptions about the data set that the model is used to describe). These conditions are designed to balance creditability (the ability to process data in a form that will be useful to researchers) with manageability (the technique's ability to simplify many mathematical derivations and operations). If valid results are to be obtained, the researcher must assume that his or her data set is similar to the type of data set required by the statistical procedure chosen.

Seldom, however, do data sets adhere perfectly to the assumptions a statistical model was developed to handle. According to Glass, Peckham and Sanders (1972), the question that the researcher must ask in reference to the data collected is not *whether* the assumptions are satisfied, but instead, *are the violations that do occur extreme enough to compromise the validity of the results?*

Box and Anderson (1955) noted that to fulfill the needs of the researcher, statistical criteria should: (a) be sensitive to change in the specific factors being tested (in other words, they should be powerful) and (b) they should be insensitive to changes in extraneous factors of a magnitude likely to occur in practice (in other words, they should be robust).

Literature Concerning the Assumptions of the Oneway, Fixed-Effects ANOVA

In 1972, Glass et al. identified three assumptions of concern for the ANOVA. The first of these is additivity--that is, each observation must be the simple sum of three components: the grand mean (μ), the effects of the treatment (a_j) and the error associated with the individual observation (e_{ij}). The presence of additivity is important because the least amount of information is lost in an additive model (Cochran, 1947). The second assumption is that the sum of the treatment effects equal zero. Glass et al. argued that this assumption is actually a mathematical restriction adopted to allow for a unique solution to the least-squares equation, rather than an assumption per se. Finally, the third assumption is that errors made while using the model should be normally distributed with a population mean of zero and a variance of 2. This third assumption involves the nature of the errors in the population from which the data originates, and takes three distinctive forms: (a) normality of the error distribution, (b) homogeneity of group variances, and (c) the independence of errors. Independence of errors is, of course, a methodological concern. Therefore it is forms (a) and (b) of the third assumption that are the subject of most theoretical and empirical research into ANOVA.

Homogeneity of Variance. This assumption was first identified in the classical 1908 paper "The Probable Error of the Mean" by The Student (Gossett); however, the publishing of empirical results in this area would wait until the work of Hsu (1938, as cited by Scheffé, 1959). Active research concerning the assumption of homogeneity of variance has continued even until today. Many of the published studies suggest that the *F* test with equal sample *n*'s is robust when faced with the single violation of the assumption of unequal group variances as long as the ratio of the largest to smallest group variances does not exceed 3 (e.g., Glass et al., 1972). Some studies (e.g., Shields, 1978) suggest that the degree of robustness present may be offset by the loss of power that is the result of using a parametric test when heterogeneity of group variances is present. The validity of the *F* ratio, however, is questionable in situations where both the sample sizes and variances are unequal. When cell sizes are unequal and two groups are involved, research suggests that inflated Type I error rates occur when the larger group size is paired with the smaller group variance (e.g., Box, 1954; Scheffé, 1959). Tomarken and Serlin (1986) have argued that ANOVA may not be the best choice in the presence of heterogeneity of variances, especially when many groups are to be compared. Their research suggested that the effects of variance heterogeneity increases as the number of groups to be compared increases. But the most surprising results of recent years, however, came in a meta-analytic study conducted by Harwell et al., (1990, 1992). They suggested that even when sample *n*'s are equal, inflated Type I errors are possible when the ratio of largest to smallest variance is as small as 2. Thus, Harwell et al. (1990) wrote, "... researchers should not rely on equal sample sizes to neutralize the effects of heterogeneous variances" (p. 23).

Normality of the Distribution of Errors. Research dating back to the 1920s has investigated violations of this assumption. Games and Lucas (1966) suggested that skewed distributions are a greater threat to robustness than leptokurtic or platykurtic distributions; however, this claim is not consistent with Pearson's 1929 power analysis among balanced designs. Assuming a distribution with a mean of 0 and variance of 1, the third moment (from which skewness is mathematically derived) is defined as follows:

$$\beta_1 = \frac{\sum (\chi - \mu)^3}{3}$$

while the fourth moment (used to calculate kurtosis) is defined as:

$$\beta_2 = \frac{\sum (\chi - \mu)^4}{4}$$

Norton (1952, cited in Glass et al., 1972) examined the degree of skewness in data distributions and found a moderately skewed distribution as having a skew value around .5, while the skew value for an extremely skewed distribution was around 1.0. A perfectly symmetrical distribution (in other words, a distribution with no skew) has a skew of 0. A perfectly mesokurtic distribution has a kurtosis of 0. Distributions with kurtosis significantly greater than 0 are leptokurtic, while those significantly less than 0 are platykurtic.

Looking at the effects of skewness in the single sample *t*-test, Pearson (1929) and Scheffé (1959) found that if the difference between the sample and population mean is positive and the distribution is positively skewed, then actual power will exceed nominal power. However, if the difference between the sample and population means is positive and the distribution is negatively skewed, then the actual power is less than nominal power. Games and Lucas (1966) suggested that *F* test results may improve when the procedure is conducted on data that has highly leptokurtic error distributions, while *F* test results for data with platykurtic error distributions tend to be adversely affected.

Extension of ANOVA Assumptions to ANCOVA. The simplest form of the analysis of covariance (which consists of one independent, one concomitant, and one dependent variable) is an extension of the oneway, fixed-effects ANOVA. According to Cochran (1957) and Winer (1962), the assumptions previously discussed in regards to ANOVA apply to ANCOVA as well, provided that the concomitant variable is normal. It is for this reason that empirical testing of either of these single violations in the ANCOVA case is scarce.

The sensitivity of the *F* test in ANCOVA to departures from normality in the dependent variable depends on the degree of nonnormality that is found in the concomitant (Potthoff, 1965). Similar results were found in Atiquallah's theoretical treatise (1964): If X (the concomitant) is a normally distributed random variable, nonnormality in the dependent variable has little effect on the *F* test. If, however, the concomitant is a random variable that is not normally distributed, then there will appear an increased sensitivity of the *F* test to nonnormality in the dependent variable.

The Seven Assumptions of the Analysis of Covariance

Elashoff (1969) and McLean (1979, 1989) reported the following seven assumptions associated with ANCOVA: (a) The cases are assigned at random to treatment conditions; (b) the covariate is measured error-free (that is, there is a perfect reliability in the measurement of the covariate); (c) the covariate is independent of the treatment effect; (d) the covariate has a high correlation with the dependent variable; (e) the regression of the dependent variable on the covariate is the same for each treatment group; (f) for each level of the covariate, the dependent variable is normally distributed; and (g) the variance of the dependent variable at each given value of the covariate is constant across treatment groups. These assumptions can be classified as falling into one of two categories: (a) assumptions that are concerned with the research design and sampling (methodological assumptions) and (b) assumptions that are concerned with the numerical form of the data set and the population from which it came (data set assumptions).

Methodological Assumptions. Two of the ANCOVA assumptions deal with the research design and sampling: (a) The cases are assigned to random treatments (randomization) and (b) the covariate has perfect reliability. Concerning the issue of randomization, Evans and Anastasio (1968) distinguished three separate situations: (a) Individuals are assigned to groups at random after which the treatments are randomly assigned to the groups; (b) intact groups are used, but treatments are randomly assigned to the groups; and (c) intact groups are used where treatments occur naturally rather than being randomly assigned by the researcher. They maintain that ANCOVA is appropriate for the first situation, can be used with caution in the second, but should be abandoned altogether (perhaps in favor of the less restraining factorial block ANOVA design) in the third. Two reasons are provided for their recommendations: First, it is never quite clear whether the covariance adjustment has removed all of the bias when proper randomization has not taken place, and second, when there are real differences among the groups, covariance adjustments may involve computational extrapolation.

A number of researchers (e.g., Loftin & Madison, 1991; McLean, 1974; Raajimakers & Pieters, 1987; Thompson, 1992) have addressed the issue of an unreliable covariate. Raajimakers and Pieters (1987) noted that there are two ways that the researcher can conceptualize covariate reliability. If one assumes that the dependent variable is linearly related to the observed value of the covariate, then the ANCOVA results will retain their statistical validity. If, on the other hand, it is assumed that the dependent variable is linearly related to the underlying true score on the covariate (rather than the sample of scores that were actually observed), then the resulting *F* ratio will produce biased results. McLean's research, however, suggested that the issue of perfect reliability becomes less of a threat to the validity of the *F* ratio if there is an independence of the covariate measure and the treatment groups.

The Covariate's Relationship with the Independent and Dependent Variables. The covariate should have no significant correlation with the independent variable, yet be highly correlated with the dependent variable. Feldt (1958) recommended the use of a covariate only when the 0-order correlation between the covariate and the dependent variable is $r \geq 0.6$. McLean (1979, 1989) saw the relationship between the covariate and the independent variable to be the most fundamental of all of the assumptions, and suggested that ANCOVA not be performed until after the data has been tested to see if it meets this assumption. If this assumption is not met, the *F* test results are not invalidated as such; however it reduces the ANCOVA's efficiency to slightly below that of doing a simple oneway ANOVA on the same data.

Homogeneity of Group Regression Slopes. This assumption requires that the slope of the regression line between the concomitant and dependent variables be the same for all levels of the grouping variable (see McLean, 1979, 1989; Thompson, 1992). The problem, if this assumption is violated, is analogous to trying to interpret main effects in the presence of significant interactions in an n-way factorial ANOVA. If heterogeneous regression slopes are suspect, the researcher would be wiser to use the randomized block ANOVA rather than ANCOVA.

Peckham (1968), McClaren (1972) and Hamilton (1972) have investigated the effects of violation of this assumption. Peckham varied regression slopes, the number of groups, and the sample size, though he limited himself to equal groups. Values of the concomitant variable were fixed and chosen to conform as closely as possible to a normal research situation. He found that there were small discrepancies in the actual vs. theoretical significance levels when the slopes were varied. He also found that as the degree of heterogeneity of the regression slopes increased, the heterogeneity of group variances likewise increased, and therefore the empirical rate of the Type I errors decreased from what is suggested by normal theory.

McClaren found similar results to Peckham when he looked at equal samples; however he extended his study to unequal groups. With the unequal group n's, McClaren found results similar to those reported by Box (1954) and Scheffé (1959); that is, when the smallest regression coefficient and the largest variance were combined with the smallest sample size, the empirical significance levels were biased in a non-conservative direction, and, likewise, when the pairings were reversed, the test became conservative.

When Hamilton (1972) conducted his study, he limited his analysis to two groups. He used the same combination of equal sample sizes, number of groups, and regression coefficients as Peckham and McClaren, yet failed to replicate their findings. Whereas Peckham and McClaren observed a conservative bias in empirical alpha levels when sample n's and regression slopes were heterogeneous, Hamilton's values were close to nominal alpha. It is unclear why there is a discrepancy in the results of the three studies (Shields, 1978). Theoretical work by Atiquallah (1964), however, suggested that ANCOVA should be robust enough to the violation of the single assumption of homogeneity of regression in situations where the sample size is large and the means of the concomitant variable(s) are equal. Otherwise, Atiquallah suggested, the test should be biased in a conservative direction.

Homogeneity of Variances and Nonnormal Error Distributions in ANCOVA. As has been discussed previously, most researchers simply accept the claim by Cochran (1957) and Winer (1962) that the effects of the simple ANOVA violations are equally viable when the model is extended to include one or more concomitant variables.

Research Methodology

Goal of the Research

This research is an exploratory study of the effects of both single and compound violations of the mathematical conditions (i.e., assumptions) underlying use of the analysis of variance and covariance designs. Monte Carlo methodology was used, allowing for the empirical investigation of problems identified by theoretical mathematicians as potential threats to the robustness of the ANOVA and/or ANCOVA results under conditions common to research practitioners in the behavioral sciences, social sciences, and education. Because of advances both in methodological techniques and computing technology, the capability has emerged to study this topic in depth, yet with a global perspective not possible just a few years ago. Capitalizing on these advances, this study has integrated into one comprehensive laboratory experiment a vast array of previously defined and substantively interrelated research avenues that have spanned across seven decades of statistical inquiry.

Specifically, this research explores the following violations that can occur in a researcher's data set: heterogeneity of group variances, skewness, non-mesokurtic distributions, and (in ANCOVA) heterogeneity of regression slopes and use of a skewed concomitant.

Information about the Computing Environment and the Programs Written to Conduct the Simulations

The statistical simulations were conducted on a Digital Equipment Corporation VAX 6430 mainframe computer with 128 M-bytes of MOS memory and 32 gigabytes of disk storage space. The simulation itself consisted of two sets of eight FORTRAN 77 programs written especially for this research: The first set of programs (phase 1 of the simulation) conducted simulations that used a normally distributed covariate vector, while the second set of programs (phase 2 of the simulation) conducted the same analyses using a skewed covariate vector. The data generated by the experiments in phase 1 were used again in phase 2 with one exception: The concomitant vectors generated for phase 1 were mathematically perturbed to produce the skewed concomitant vectors needed for phase 2.

The Simulation Process, Part I: Within a Single Replication of an Experiment

Four experimental situations were simulated in each of the two phases of the simulation: three balanced designs (i.e., equal sample sizes) and one unbalanced design (i.e., unequal sample sizes). For explanation purposes, these four experimental situations will be referred to in this text as experiments A, B, C, and D. Experiment A tested the ANOVA and ANCOVA *F* statistic when three equal groups of size 15 were used. Experiment B involved simulation using three equal groups of size 30, while experiment C tested the *F* statistic when three equal groups of size 45 were used. The fourth condition, experiment D, involved simulation of the ANOVA and ANCOVA *F* statistic when three unequal sized groups (n's = 15, 30, and 45) were used.

Experiments A, B, and C of phase 1 were used to generate the data. Experiment D, on the other hand, did not generate data. Instead, it imported grouping, concomitant, and dependent vectors from the data generating experiments, so that the first group had a size of 15, the second group 30, and the third group 45. The use of data in experiment D which was not independent of the data used in experiments A, B, and C was to facilitate the comparison of the balanced and unbalanced design results. By using the same data, a major source of sampling error was eliminated, sampling error that otherwise might confound interpretation of the results. Likewise, phase 2 of the study (for both the balanced and unbalanced designs) imported data that was created in the data generating experiments of phase 1 with only one change: The concomitant vectors, which were normally distributed when they were originally created in phase 1, were perturbed to create moderately skewed covariates.

The data generated for experiments A, B, and C were created using the International Mathematical and Statistical Libraries (IMSL) subroutine RNVMN, a subroutine which is designed to create multivariate normal distributions with means equal to 0, standard deviations equal to 1, and correlations between vectors that can be specified beforehand by the user. Data for each treatment level were created separately using IMSL. This made it possible to obtain the unequal group regression slopes desired for the second concomitant vector. For the first concomitant vector, the correlation between all groups and the IMSL created dependent variable was set at $r = 0.707$, thus simulating homogeneity of regression slopes. For the second concomitant, heterogeneity of regression slopes was simulated by having IMSL create concomitant vectors for group 1 that had a correlation of $r = 0.6$ with group 1's dependent vector, a correlation of $r = 0.707$ between the group 2 concomitant and dependent vectors, and $r = 0.8$ between the third group's concomitant and dependent vectors.

The next step of the data creation process would require that duplicate copies of the dependent vector be created and then perturbed in a systematic fashion to simulate specific skew and/or kurtotic conditions. Therefore, it was imperative that the originally created vectors themselves have the purported mean, variance, skewness, and kurtosis. This was accomplished by building a testing procedure into the data generating FORTRAN programs.

By using this testing procedure, dependent vectors created by IMSL were tested to see if their skew and kurtotic values fell within the 95% confidence bands that surround zero skew and kurtosis for the specific group size. Therefore, for experiment A (where the group size was 15), all dependent vectors generated by IMSL were tested to determine if their skew was between -1.137 and

1.137, while the kurtosis was tested to see if it fell between -4.038 and 4.038. If either value was not within these limits, then the data created by IMSL was discarded, and new data created and tested. Likewise for experiment B (n = 30), skew values were tested to assure that they fell between the values of -0.837 and 0.837, while kurtosis values were checked to assure that they were between -3.478 and 3.478. For experiment C (n = 45), confidence bands for skew were -0.693 and 0.693, while they were -3.205 and 3.205 for kurtosis. For all of the data generating experiments, the data were retained only when both the skew and kurtosis values of the dependent vectors created by IMSL fell within these limits.

These checks assured that the base vectors (that is, those created originally by IMSL) were normally distributed, with no significant skew or kurtosis. This, in turn, allowed for mathematically valid perturbations to be performed on them. The checks do, however, represent a departure from the sampling procedure characteristic of more traditional Monte Carlo studies. Using the more traditional approach, parent populations with the desired mathematical characteristics are created. Out of these parent populations, repeated samples of the desired size are randomly selected and tested. While this methodology is more generalizable because of its ability to simulate the central limit theorem, it also allows the inclusion of samples with skew and/or kurtosis radically different from what they are purported to be. Therefore, when differences between the empirical results and normal theory surface, it is unclear to what degree these differences are the result of the known mathematical characteristics of the parent population, and at what point they become the result of selected samples that, as the result of pure chance, possess mathematical characteristics far different from their parent population.

After IMSL created acceptable concomitant and dependent vectors, phase 1 of the simulation required that the normal dependent vector be duplicated, then algebraically perturbed to simulate 27 different mathematical conditions. Distortions of distributional shape were imposed on the data first. This was done using Fleishman's method (1978), which uses the following function:

$$Y = a + bX + cX^2 + dX^3$$

where the coefficients b, c, and d are obtained by consulting a special table compiled by Fleishman (1978), and the coefficient a has the same absolute value as the coefficient c, but the opposite sign. Using this polynomial expression, the base dependent vector's values were substituted for X, while the resulting Y values formed a distribution with the desired shape.

Use of Fleishman's (1978) function allowed the desired combination of skew and kurtosis values to be created within a tolerable margin of error without distorting the original mean or standard deviation. The originally created (i.e., base) dependent vector was normal, with no skew and kurtosis. After Fleishman's (1978) formula was imposed on duplicate copies of the original dependent vector, the following combinations of skew and kurtosis were simulated: moderately skew (skew = 0.5, kurtosis = 0), platykurtic (skew = 0, kurtosis = -0.5), leptokurtic (skew = 0, kurtosis = 2), moderately skewed and platykurtic (skew = 0.5, kurtosis = -0.5), moderately skewed and leptokurtic (skew = 0.5, kurtosis = 2), and extremely skewed and leptokurtic (skew = 1, kurtosis = 2). This allowed for every combination of skew and kurtosis with two exceptions: an extremely skewed and platykurtic distribution and an extremely skewed and mesokurtic distribution. Neither of these shapes was possible to obtain using the coefficients published by Fleishman (1978).

After the algebraic manipulations to distort shape, seven dependent vectors possessing the characteristics described above were available. Each of these seven vectors was then duplicated three more times, and the three duplicate vectors for each shape linearly transformed. After the duplicate vectors were transformed, there were four different group variance ratios for each of the seven distributional shapes: 1:1:1 (homogeneity of variance), 1:1.5:2 (slight heterogeneity of variance), 1:2:3 (moderate heterogeneity of variance) and 1:3:5 (extreme heterogeneity of variance). These inter-group variance conditions were chosen specifically to allow the testing of Harwell et al.'s 1990 claim (that differences from normal theory may be present in balanced designs when the ratio between the largest and smallest variance is 2) against the standard set by Glass et al. in 1972 (that differences from normal theory do not emerge in balanced designs until the ratio between the largest and smallest variance is at least 3).

As has been mentioned previously, no new data was generated for experiment D (the unbalanced design). Instead, a systematic process imported vectors already created. Specifically, treatment level (group) 1 from experiment A, group 2 from experiment B, and group 3 from experiment C were imported. This created the unequal n simulation where group 1 had an n = 15, group 2 had an n = 30, and group 3 had an n = 45.

Therefore, in the end 28 different dependent vectors, two concomitant vectors, and a grouping vector were either created for or imported into each replication of all of the experiments. For the ANOVA simulations, the grouping vector was combined with each of the dependent vectors, computing 28 *F* ratios (one for each combination of skew, kurtosis, and variance). For the ANCOVA simulations of phase 1, the first concomitant vector was combined with the grouping vector and each of the 28 dependent vectors to calculate 28 ANCOVA *F* statistics using a normal covariate with equal regression slopes. The second concomitant vector was then combined with the grouping vector and each dependent vector to calculate 28 ANCOVA *F* statistics using a normal covariate with unequal regression slopes.

As has been mentioned before, the experiments of phase 2 used the same data that was created in phase 1; however the normal covariate created in phase 1 was skewed by Fleishman's function (skew value = 0.75). Phase 2 was designed to test ANCOVA when the only difference was use of a skewed concomitant rather than a normal one. Therefore, only 56 additional *F* statistics were calculated per replication in this phase: 28 involving use of a skewed covariate and equal slopes and 28 involving use of a skewed covariate and unequal slopes.

Besides using IMSL subroutines to generate the data, IMSL subroutines were also incorporated into the FORTRAN programs to calculate the *F* ratios. Specifically, IMSL subroutine AONEW was used to obtain the ANOVA *F* values, while subroutine AONEC was used to calculate the ANCOVA *F* values.

The Simulation Process, Part II: The Global Design

As has been mentioned previously, phase 1 of the study was designed to test the ANOVA and ANCOVA *F* test when a normal covariate was combined with violations of one or more of the following assumptions: normal skew, normal kurtosis, homogeneity of variances, and (in the ANCOVA) situation, homogeneity of regression slopes. Phase 2 of the study conducted the same analyses using a skewed covariate rather than a normal one.

Glass et al. (1972) recommended that the sampling distributions created in Monte Carlo studies have a minimum of 2,000 *F* ratios each. For the three experimental conditions involving equal group n's, sampling distributions of 4,000 (twice the minimum recommended by Glass et al.) were created. In the experimental condition involving unequal n's and homogeneity of variances, *F* sampling distributions of 4,000 *F* ratios were also created. In the situation where unequal n's were combined with heterogeneity of variances, however, the combination of variance ratios and group sizes were varied so that two sets of sampling distributions with 2,000 *F* ratios each were developed: one set where the largest group variance was combined with the largest sample size and the other set where the largest group variance was combined with the smallest sample size. This was done since previous literature suggests that heterogeneity of group variances produces different effects in the unequal n situation, depending on the combination of sample size and magnitude of group variances (e.g., Box, 1954; McClaren, 1972; Scheffé, 1959). The relationship between sample size and group regression coefficients was fixed for those analyses that involved unequal group slopes, therefore the process of varying the magnitude of group variances with the sample size produced the following triple combinations for analysis in the ANCOVA simulations: (a) the largest group size with largest group variance and largest regression coefficient, and (b) the largest group size with the smallest group variance and largest regression coefficient. Previous literature (e.g., Glass et al., 1972; Shields, 1978) suggests that the additivity of effects should produce dramatic differences in these two combinations.

After running all four sets of experiments in both phases of the simulation, a total of 420 empirical sampling distributions of 4,000 *F* ratios each were created, representing all single and compound data set violations for the balanced ANOVA and ANCOVA simulations. Another 35 sampling distributions of 4,000 *F* ratios each included all unbalanced ANOVA and ANCOVA simulations with homogeneity of group variances. Finally, another 210 empirical sampling distributions of 2,000 *F* ratios each were created, representing all single and compound data set violations having both heterogeneous variances and unequal sizes.

Of these 665 *F* sampling distributions, four ANOVA and four ANCOVA *F* distributions were created using data that did not contain any violation under study. These eight sampling distributions (one ANOVA and one ANCOVA for each of the four experimental conditions A, B, C, and D) served as a baseline against which other distributions could be compared, and served as a check to make sure that the simulation was operating properly.

Statistical Analysis of the Sampling Distributions

In addition to qualitative evaluation of the sampling distributions, statistical analysis of the data was performed using the Kolmogorov-Smirnov one sample test at the $p < .05$ and (where applicable) $p < .01$ levels of significance. The non-parametric test was employed to compare the empirical sampling distributions with the appropriate theoretical (i.e., nominal) *F* distribution at four key points in the nominal *F* tail region: .90, .95, .975, and .99. These points, of course, are the points on the nominal F curve used by practitioners when testing for significance at the $p < .10$, $p < .05$, $p < .025$, and $p < .01$ levels of significance respectively. In addition, the means, standard deviations, skew, and kurtosis values for each of the entire populations of data generated in the study were calculated and inspected to assure the integrity of the results.

Results

Summarized here are the specific results of the effects of violations of data set assumptions for the analysis of variance and covariance statistical models. Since the integrity of the results is dependent on the quality of the data produced, the first section will discuss the descriptive statistics for the entire population of data produced for this simulation. The second section will summarize the effects of violations in the ANOVA situation. The third and fourth sections will summarize the effects of violations on the ANCOVA.

Analysis of the Population Data

All data created in each of the replications of the data generating experiments were retained in order to verify the integrity of the results. In the actual process of creating the data, the vectors for each treatment level were created individually, then merged with the vectors for the other treatment levels before ANOVA or ANCOVA could be performed. The population vectors were checked for each treatment level separately; then the full vectors (which consisted of the three treatment levels merged together) were also checked. All population vectors, including the base vectors created by IMSL and the vectors perturbed by use of Fleishman's function, were at or very near their target parameters.

The size of the populations are worth noting. In their classic 1972 paper, Glass et al. suggest that populations with the desired characteristics have a minimum of 10,000 points each. The population *N*'s used in this study were considerably larger than the minimum standard: 180,000 for each of the full population vectors created in experiment A, 360,000 for the full population vectors created in experiment B, and 540,000 for the full population vectors created in experiment C. The population statistics for the vectors created to simulate heterogeneous variances were also checked. As expected, the simple linear multiplication that changed their variances did not change the vectors means, skew, or kurtosis.

Effects of Data Set Assumptions on the Analysis of Variance

For all of the analyses to follow, comparisons were made between the empirical F sampling distributions and the theoretical (i.e., nominal) F distributions expected using normal theory. The values included on both of the tables were calculated by subtracting the number of F ratios for each mathematical condition expected to be nonsignificant under normal theory from the number which actually were observed to be nonsignificant in this simulation. There were 4,000 F ratios in each of the F distributions for those designs which were balanced as well as for those unbalanced designs which had homogeneity of group variance. Therefore, for these cases a given mathematical condition was found to produce results significantly different from normal theory when the subtraction found a difference greater than or equal to ±77 ($p < .05$) and greater than or equal to ±96 ($p < .01$). For those unbalanced designs having heterogeneity of variances, however, there were only 2,000 F ratios in each of the F distributions. Therefore, for these mathematical conditions, significant differences were obtained when the subtraction found a difference greater than or equal to ±55 ($p < .05$) and greater than or equal to ±68 ($p < .01$).

When the group size was 15 and all groups were equal, no empirical sampling distribution was found to have Type I error rates significantly different from what would be expected under normal theory, although the sampling distribution that was based on an extremely skewed and leptokurtic dependent vector with extreme heterogeneous variances (variance ratio 1:3:5) came within one F value of being significant at the $p < .05$ level. When violations were imposed on the dependent vectors with groups of size 30 and 45, no empirical distributions were found significantly different from the nominal F distribution at the $p < .05$ level (see Table 1).

For the equal n experiments, the differences between the empirical and theoretical F sampling distributions were largest when the sample size was small and became smaller as the group sizes grew larger. It is possible that this trend, found in the ANCOVA results as well, may be due to the fact that confidence bands increase when sample size is small. All dependent base vectors created by IMSL, as one will recall, were tested to exclude extreme vectors with mathematical characteristics different from those purported. It is possible that when sample sizes are less than 30, confidence bands are not narrow enough to eliminate all samples that are not representative of their parent populations.

Table 1
Maximum Differences Between the Empirical and Nominal Sampling Distributions for the ANOVA and ANCOVA Simulations

Largest to Smallest	BALANCED DESIGNS												UNBALANCED DESIGNS						
Group Variance Ratios	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	1:1	1:2	1:3	1:5	2:1	3:1	5:1
ANOVA																			
	Sample Size = 15				Sample Size = 30				Sample Size = 45					Sample Sizes = 15, 30, 45					
Distributional Shape																			
Normal	17	-15	-34	-60	17	-17	-24	-37	-9	-16	-17	-34	27	88[b]	107[b]	115[b]	-87[b]	-136[b]	-194[b]
Platykurtic	30	-8	-33	-63	21	-14	-23	-39	6	-12	-17	-22	26	87[b]	109[b]	119[b]	-86[b]	-142[b]	-195[b]
Leptokurtic	24	-11	-27	-52	-8	-28	-31	-44	-14	-11	-20	-35	16	85[b]	102[b]	113[b]	-84[b]	-129[b]	-193[b]
Moderate Skew	24	-13	-39	-73	9	-17	-28	-44	11	-11	-19	-31	28	83[b]	103[b]	112[b]	-86[b]	-144[b]	-195[b]
Mod. Skew & Platy.	27	-16	-39	-67	14	-10	-18	-35	11	-7	-13	-19	34	91[b]	111[b]	118[b]	-84[b]	-132[b]	-201[b]
Mod. Skew & Lepto.	15	-14	-27	-65	-6	-24	-34	-43	-16	-14	-20	-32	19	84[b]	99[b]	111[b]	92[b]	-136[b]	-198[b]
Extreme Skew & Lepto.	23	-17	-43	-85	2	-12	-30	-46	-16	-20	-25	-44	24	87[b]	106[b]	115[b]	-96[b]	-143[b]	-200[b]
ANCOVA: Normally Distributed Covariate with Equal Regression Slopes																			
	Sample Size = 15				Sample Size = 30				Sample Size = 45					Sample Sizes = 15, 30, 45					
Distributional Shape																			
Normal	9	-23	-49	-70	49	27	14	-7	11	16	-13	-26	31	67[a]	83[b]	87[b]	-86[b]	-142[b]	-186[b]
Platykurtic	23	-16	-50	-65	51	37	15	-6	19	16	19	-25	37	72[a]	84[b]	90[b]	-80[b]	-140[b]	-181[b]
Leptokurtic	13	-18	-33	-63	11	11	6	-14	8	-6	-10	-23	22	61[a]	75[b]	78[b]	-90[b]	-129[b]	-177[b]
Moderate Skew	-19	-21	-48	-81	40	31	12	-11	20	20	18	-22	15	69[a]	97[b]	90[b]	-134[b]	-221[b]	-325[b]
Mod. Skew & Platy.	8	-21	-49	-80	38	18	15	-12	16	12	11	-21	46	77[b]	97[b]	97[b]	-84[b]	-159[b]	-197[b]
Mod. Skew & Lepto.	13	-17	-42	-69	16	11	7	-25	-4	8	-14	-24	24	79[b]	92[b]	75[b]	-92[b]	-140[b]	-190[b]
Extreme Skew & Lepto.	11	-24	-58	-95[a]	24	11	-13	-37	17	-11	-13	-25	32	50	68[a]	75[b]	-96[b]	-140[b]	-207[b]
ANCOVA: Normally Distributed Covariate with Unequal Regression Slopes																			
Distributional Shape																			
Normal	-23	-47	-67	-81	-29	-33	-36	-52	22	30	21	8	-191[b]	-9	19	31	-182[b]	-221[b]	-261[b]
Platykurtic	-26	-42	-62	-79	-24	-29	-33	-42	30	32	32	23	-179[b]	-14	24	34	-170[b]	-206[b]	-253[b]
Leptokurtic	-20	-40	-52	-71	-37	-41	-49	-51	24	10	-10	-15	-177[b]	-11	22	44	-182[b]	-220[b]	-282[b]
Moderate Skew	-26	-44	-60	-84	-20	-30	-39	-41	25	29	18	-13	-165[b]	-8	30	41	-171[b]	-229[b]	-267[b]
Mod. Skew & Platy.	-21	-36	-53	-71	-14	-32	-34	-35	20	40	29	15	-141[b]	9	24	51	-160[b]	-216[b]	-251[b]
Mod. Skew & Lepto.	-30	-40	-56	-70	-22	-30	-38	-40	24	16	11	-13	-165[b]	-5	23	48	-176[b]	-222[b]	-278[b]
Extreme Skew & Lepto.	-32	-35	-64	-102[a]	-27	-26	-35	-51	27	20	13	14	-170[b]	7	30	44	-168[b]	-208[b]	-271[b]
ANCOVA: Skewed Covariate with Equal Regression Slopes																			
	Sample Size = 15				Sample Size = 30				Sample Size = 45					Sample Sizes = 15, 30, 45					
Distributional Shape																			
Normal	±4	13	-18	-28	43	28	18	-17	17	7	-9	-21	27	83[b]	100[b]	103[b]	-106[b]	-152[b]	-194[b]
Platykurtic	8	16	-22	-27	48	28	23	-11	17	7	3	-17	30	87[b]	94[b]	102[b]	-97[b]	-194[b]	-193[b]
Leptokurtic	11	-2	-12	-20	22	14	8	-13	18	±6	-8	-19	18	72[a]	85[b]	93[b]	-98[b]	-142[b]	-190[b]
Moderate Skew	6	-9	-25	-39	25	25	9	-19	9	12	-18	-21	30	83[b]	98[b]	101[b]	-99[b]	-151[b]	-290[b]
Mod. Skew & Platy.	7	-15	-27	-45	42	28	32	-10	17	8	-9	-19	55	88[b]	101[b]	103[b]	-100[b]	-198[b]	-203[b]
Mod. Skew & Lepto.	13	-9	-13	-21	23	7	-8	-12	5	15	-8	-20	14	68[a]	85[b]	98[b]	-94[b]	-140[b]	-192[b]
Extreme Skew & Lepto.	7	-15	-17	-47	18	-10	-14	-28	15	10	-15	-24	30	78[b]	90[b]	94[b]	-95[b]	-141[b]	-191[b]
ANCOVA: Skewed Covariate with Unequal Regression Slopes																			
Distributional Shape																			
Normal	-17	-21	-26	-40	-25	-27	-35	-45	14	9	10	-8	-166[b]	-5	21	46	-169[b]	-200[b]	-249[b]
Platykurtic	-13	-21	-22	-29	-18	-27	-30	-38	13	16	19	-7	-170[b]	-10	31	31	-163[b]	-200[b]	-249[b]
Leptokurtic	-13	-16	-24	-31	-25	-34	-35	-45	17	9	-4	-18	-154[b]	-6	31	50	-154[b]	-206[b]	-249[b]
Moderate Skew	-28	-24	-34	-49	-22	-28	-33	-42	19	11	3	-8	-158[b]	-8	15	33	-152[b]	-202[b]	-255[b]
Mod. Skew & Platy.	-18	-22	-26	-49	-13	-26	-33	-38	27	9	10	18	-166[b]	-5	25	35	-158[b]	-210[b]	-245[b]
Mod. Skew & Lepto.	-23	-33	-34	-38	-24	-37	-38	-43	21	11	7	-14	-162[b]	-8	23	39	-164[b]	-218[b]	-262[b]
Extreme Skew & Lepto.	-34	-29	-41	-55	-30	-34	-43	-52	21	11	7	-16	-162[b]	10	21	37	-175[b]	-219[b]	-267[b]

[a] Significant at the $p < .05$ level.
[b] Significant at the $p < .01$ level.

No significant differences were found in the unbalanced designs having homogeneous variances. Significant differences did emerge, however, when the unbalanced ANOVA was combined with even the slightest degree of heterogeneity of variance (group variances as small as 1:1.5:2). Further analysis revealed two different trends, depending on whether the largest variance was coupled with the largest or smallest group. Specifically, when the smallest group had the largest variance, all empirical sampling distributions were significantly less than the theoretical *F* distribution at the $p < .01$ level of significance. When the largest group contained the largest variance, however, the opposite trend developed: Sampling distributions having heterogeneity of variances were found to be significantly greater than theoretical *F* at the $p < .01$ level.

Effects of Assumption Violations on the Analysis of Covariance Using a Normal Concomitant

Differences between the empirical and nominal *F* sampling distributions for the ANCOVA simulations using a normal covariate are found in Table 1 also. For the balanced design using small but equal group sizes ($n = 15$), the only compound violation that had a significant impact on the resulting empirical sampling distribution was the combination of an extremely skewed and leptokurtic shape with extreme heterogeneity of variances (ratio of 1:3:5), which was significant at the $p < .05$ level. In those simulations that had equal n's of size 30, no significant differences emerged. Equal n's of 45 showed more of the same; no significant differences were found even when extreme heterogeneity of variances was combined with unequal regression slopes.

Among the unbalanced ANCOVA simulations involving homogeneity of variances, no significant differences emerged as long as the regression slopes were equal. When the group slopes were unequal, however, all analyses were significant at the $p < .01$ level.

In those ANCOVA simulations involving both equal slopes and heterogeneous variances, significant differences emerged--most at the $p < .01$ level. Different trends emerged, however, depending on whether the largest variance was in the largest or smallest group. When the largest variance was found in the largest group, the number of Type I errors was significantly higher than what was expected under normal theory. When the largest variance was found in the smallest group, however, the number of Type I errors was significantly less than what would be expected under normal theory.

When unequal group slopes were coupled with heterogeneous variances, a different pattern emerged. When the largest variance was found in the smallest group, significant differences (at the $p < .01$ level) emerged, raw differences that were much higher than when the largest variance was paired with the smallest group in the equal n simulation. When the largest variance was paired with the largest group size, however, no significant differences could be found. It should be mentioned at this point that the largest group correlation (slope) is found in the third treatment group for both of these situations. Apparently, the coupling of the largest variance with the largest group size and largest regression slope improves the fit between the empirical and theoretical sampling distributions, while the coupling of the largest variance with the smallest group size and the smallest regression slope increases the disparity between the empirical and theoretical sampling distributions.

Effects of Assumption Violations on the Analysis of Covariance Using a Skewed Concomitant

Differences between the empirical and nominal *F* sampling distributions for the ANCOVA simulations using a skewed covariate are also found in Table 1. For balanced designs involving small groups ($n = 15$) and a skewed covariate, no significant differences emerged. In fact, those (statistically nonsignificant) differences that did emerge tended to be smaller in magnitude than those found when the same dependent vectors were used with normal covariates. The same can be said for the balanced designs using groups of 30 and 45.

When the unbalanced design was coupled with equal slopes and homogeneity of variances, no significant differences emerged. When the unbalanced design was coupled with heterogeneous slopes and homogeneity of variances, however, differences significant at the $p < .01$ level did emerge.

When heterogeneity of variance was coupled with equal regression slopes and unequal group sizes, the patterns identified originally with use of a normal covariate emerged again. Significantly less Type I errors emerged when the largest variance was found in the largest group. However, when the largest variance was paired with the smallest group, there was a significant increase in the number of Type I errors made.

When heterogeneity of variances was coupled with heterogeneous slopes and unequal n's, patterns emerged which were similar to those identified when the normal covariate was coupled with unequal slopes. When the largest variance was found in the smallest group, significant differences (at the $p < .01$ level) emerged; raw differences which were much higher than when the largest variance was paired with the smallest group in the equal slope situation. When the largest variance was paired with the largest group size, however, no significant differences could be found. Again here, like the analyses involving a normal covariate, the smallest correlation coefficient was found in the group with the smallest size. And again the coupling of the largest variance with the largest group size and largest regression slope improves the fit between the empirical and theoretical sampling distributions, while the coupling of the largest variance with the smallest size and the smallest slope increases the disparity. Once again, it is interesting to note that in many cases, use of the skewed covariate seemed to improve the fit between the empirical and nominal F sampling distributions.

Findings and Conclusions

Balanced Designs

Previous research (Glass et al., 1972; Harwell, Hays, Olds, & Rubinstein, 1990, 1992; etc.) suggests that heterogeneity of variances is the greatest single threat to robustness. Conventional thought suggests that when a balanced ANOVA or ANCOVA is used problems arise only when the ratio of largest to smallest group variance exceeds 3. Meta-analytic findings by Harwell et al., however, suggest differently: Balanced designs may suffer from inflated Type I error rates when the ratio is as small as 2.

The group variance ratios used in this simulation were chosen to directly compare Harwell et al.'s claim against the standard set by Glass et al. (1972). No support was found for Harwell's claim; quite the contrary, there were almost no significant differences found in any of the balanced designs, even when the ratio between the largest and smallest group variance was as high as 5.

The results of this simulation when using a balanced design ANOVA or ANCOVA suggest a robustness far beyond that suggested by Glass et al. (1972). The unique methodology employed in this study may help to explain why. As part of the data generating process, the base vectors that had skew or kurtosis values significantly different from 0 were systematically discarded and new ones created. This procedure reduced the probability that the perturbations were a shape different than purported. Following removal of this sampling noise, the causes for the differences that remain are easier to isolate and interpret. Most of the studies that Glass et al. (1972) reviewed, however, used a methodology whereby parent populations with the desired characteristics were created and repeated random samples were drawn. No check was made to insure that the samples drawn possessed the mathematical properties being tested. Therefore, when significant differences emerged between the empirical and theoretical F distributions, it was unclear to what degree the differences were the result of the known mathematical characteristics and at what point they became the product of selected samples that, by the luck of the draw, possess mathematical properties far different from their parent populations.

The fact that the few significant differences that did arise in the balanced designs did so among the small group size ($n = 15$) is also worth noting. The confidence bands, used to screen out samples with mathematical characteristics different from those to be tested, are widest when the sample size is small. It is possible that some samples which should have been discarded were not because of the wide confidence bands. If this is the case, then the origin of the significant differences that emerged in the small sample size simulations remains unclear: Are they the result of violations of the assumptions under test, or are they the result of inclusion of extreme samples with mathematical characteristics different from those being tested?

Games and Lucas (1966) suggested that a skewed dependent is a greater threat to robustness than a leptokurtic or platykurtic dependent variable. Additionally, they have suggested that the validity of the F test improves for leptokurtic distributions but suffers when using platykurtic distributions. Distributional shape, however, did not prove to be a major factor in influencing Type I error rates in this simulation.

Potthoff (1965) suggests that a non-normal concomitant increases the sensitivity of F to departures from normality in the dependent variable. This research found just the opposite: The small (but statistically nonsignificant) differences that did emerge found analyses using the normal covariate--not the skewed--to be most sensitive to distortions in the dependent variable.

Unbalanced Designs

Whereas the balanced design turned out to be very robust, the same cannot be said of the unbalanced design. Statistically significant differences emerged in face of almost all conditions except some that involved only perturbations of shape. Previous research (e.g., Scheffe, 1959; Shields, 1978) have suggested that when heterogeneity of variance is coupled with unequal n's, the effect of the violation of equal variances will differ in nature depending on whether the larger group is paired with the larger or smaller variance. Specifically, they suggest that inflated Type I error rates result when there is an inverse relationship between the group size and its variance, while deflated Type I error rates will result when the larger group is paired with the larger variance.

Glass et al. (1972) suggest that the effects of nonnormal shapes and heterogeneous variances appear to be additive, something that this research supports. The idea of additive effects seems to extend beyond the match between distributional shape and heterogeneous variances, however. For instance, in the unequal n situation the smallest regression slope is paired with the smallest group size for all analyses. When this combination (which should increase the number of Type I errors made) occurs jointly with heterogeneous variances where the smallest variance is found in the smallest group (which should decrease the number of Type I errors), the net effect is a wash out; that is, no significant differences remain. Conversely, when the combination of the smallest slope and group size was paired with the largest variance, the number of Type I errors increased dramatically--higher than either one of the violating conditions alone could have produced.

Concluding Remarks

In summary, for balanced designs the ANOVA and ANCOVA *F* statistics were found to be remarkably robust when faced with most of the violations included in this simulation. The degree to which the *F* test was robust, however, was surprising. The procedure remained robust even when the ratio of largest to smallest variance was as high as 5. After the systematic removal of sampling noise due to the chance creation of skewed and/or kurtotic base vectors, *F* was found to be far more robust than previously believed. This research, however, reaffirms once again the findings of many previous studies that suggest that ANOVA and ANCOVA be avoided when group sizes are not equal.

In terms of specific recommendations to research practitioners using balanced designs, the ratio of largest to smallest group variance should continue to be checked. If the ratio is less than 3, then the researcher need not fear invalid results due to any of the data set violations included here. If the ratio is between 3 and 5, however, the researcher should test to see if his or her dependent data is within the 95% confidence bands surrounding zero skew and kurtosis. If the dependent's skew and kurtosis values are within this range, then the *F* statistic should still be sufficiently robust. If, however, either the skew or kurtotic values fall outside of the 95% confidence band, then the researcher should consider the use of a statistical procedure with less stringent assumptions.

In terms of the direction of future research, several questions remain unanswered concerning the specific findings of this simulation. First, if the balanced designs (for group n's of 30 and above) are sufficiently robust when the largest to smallest group variance ratio is as high as 5, then how high can that ratio get before robustness is significantly affected? Second, for equal sized samples smaller than size 30, are the confidence bands sufficiently narrow to provide researchers with the reassurance they need to use ANOVA or ANCOVA when the ratio of largest to smallest variance is between 3 and 5? Can use of smaller confidence bands (90% or 80% perhaps?) make up for the smaller sample size? Finally, this research used extremely unequal group sizes in the unbalanced designs (a difference of 300% between the largest and smallest groups). What would happen if the difference between the largest and smallest groups was smaller? How different can group sizes become before the robustness of the *F* statistic is jeopardized?

Finally, it should be noted that this research deals only with robustness. Robustness, however, is only the first of two issues that a researcher must consider when choosing a statistical procedure to analyze his or her data. The second issue involves power, and ultimately reduces to the following question first suggested in 1959 by Scheffe: Which procedure from among those available will produce the most statistically accurate results in a specific research situation? It is in this direction that future Monte Carlo research of this genre must direct its attention.

References

Atiquallah, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika*, *51*, 365-372.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I: Effect of inequality of variance in the oneway classification. *Annals of Mathematical Statistics, 25*, 290-302.

Box, G. E. P., & Anderson, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, 17*, 1-26.

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics, 3*, 22-38.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13*, 261-281.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-443.

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal, 6*, 383-401.

Elmore, P. B., & Woehlke, P. L. (1988). Statistical methods employed in the *American Educational Research Journal, Educational Researcher*, and *Review of Educational Research* from 1978 to 1987. *Educational Researcher, 17*(9), 19-20.

Evans, S. H., & Anastasio, E. J. (1968). Misuse of analysis of covariance when treatment effects and covariate are confounded. *Psychological Bulletin, 69*, 225-234.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika, 23*, 335-353.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

Games, P. A., & Lucas, P. A. (1966). Power of the analysis of variance of independent groups on non-normal and normally transformed data. *Educational and Psychological Measurement, 26*, 311-327.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*(3), 237-288.

Goodwin, L. D., & Goodwin, W. L. (1988). Statistical techniques in AEJR articles, 1979-1983: The preparation of graduate students to read the educational research literature. *Educational Researcher, 14*(2), 5-11.

Halpin, G., & Halpin, G. (1988, November). *Evaluation of research and statistical methodologies.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY.

Hamilton, B. L. (1972). *A Monte Carlo comparison of parametric and nonparametric uses of a concomitant variable.* Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Hammersley, J. M., & Handscomb, D. C. (1967). Monte Carlo methods. London: Methuen.

Harwell, M. R., Hayes, W. S., Olds, C. C., & Rubinstein, E. N. (1990). *Summarizing Monte Carlo results in methodological research: The oneway, fixed-effects ANOVA case.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Harwell, M. R., Hayes, W. S., Olds, C. C., & Rubinstein, E. N. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed-effects ANOVA case. *Journal of Educational Statistics, 17*(4), 315-339.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410-416.

Loftin, L., & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 133-147). Greenwich, CT: JAI Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McClaren, V. R. (1972). *An investigation of the effects of violating the assumption of homogeneity of regression slopes in the analysis of covariance model upon the F-statistic.* Unpublished doctoral dissertation, North State Texas University, Denton.

McLean, J. E. (1974). *An empirical examination of analysis of covariance with and without Porter's adjustment for a fallible covariate.* Unpublished doctoral dissertation, University of Florida, Gainesville.

McLean, J. E. (1979, November). *The care and feeding of ANCOVA.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR.

McLean, J. E. (1989, November). *ANCOVA: A review, update, and extension.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR.

Pearson, E. S. (1929). The distribution of frequency constants in small samples from nonnormal symmetrical and skew populations. *Biometrika, 19*, 151-164.

Peckham, P. D. (1968). *An investigation of the effects of non-homogeneity of regression slopes upon analysis of covariance.* Unpublished doctoral dissertation, University of Colorado, Boulder.

Potthoff, R. F. (1965). Some Scheffé type tests for some Behrens-Fisher type regression problems. *Journal of the American Statistical Association, 60,* 1163-1190.

Raaijmakkers, J. G., & Pieters, J. P. M. (1987). Measurement error and ANCOVA: Functional and structural relationships. *Psychometrika, 52*(4), 521-538.

Scheffé, H. (1959). *The analysis of variance.* New York: John Wiley & Sons.

Shields, J. L. (1978, July). *An empirical investigation of the effect of heteroscedascity and heterogeneity of variance on the analysis of covariance and the Johnson-Neyman technique.* Army Technical Paper 292, Project no. 2Q762722A777.

Student (W.S. Gossett). (1908). The probable error of the mean. *Biometrika, 6,* 1-25.

Thompson, B. (1992). Misuse of ANCOVA and related "statistical control" procedures. *Reading Psychology, 13,* iii-xviii.

Tomarken, A., & Serlin, R. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99,* 90-99.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Copyright 1994 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1994, Vol. 1, No. 2, 77-84

Effects of Item Parameters on Ability Estimation in Item Response Theory

Jwa K. Kim
Middle Tennessee State University

Effects of item parameters on ability estimation were investigated through Monte Carlo studies utilizing the Expected-A-Posteriori (EAP) estimation. The three-parameter logistic (3-PL) model was applied to all 12 situations resulting from the combination of three values of the item discriminating parameter (a = .5, 1.0, and 2.0) and four different distributions of the item difficulty parameter (Difficult, Easy, Normal, and Uniform Test). The result showed a significant effect of item discriminating parameter on standard error of ability estimation. As the discriminating parameter increases, the standard error decreases.

Estimating examinees' true ability has been the major task of testing theory. According to classical test theory introduced by Spearman (1904), the true ability (τ_{ij}) of person i for item j is the expected value of observed scores, $E(X_{ij})$. This model is a tautology, therefore cannot be tested (Lord, 1980). The model, however, has serious practical problems: test-dependent person parameter, sample-dependent item or test parameter, and the requirement of parallel tests. The first problem is that the person parameter, or the true score, in classical test theory is dependent on the item and test difficulty. A person's true score will be high if the person takes relatively easy tests, and the same person's true score will be low with relatively difficult tests. The second problem is related to the first. The item and test parameters in classical test theory, such as item *p*-value, item-test correlation, and test validity, depend on the selection of examinees. For example, the item *p*-value, the proportion of examinees who answer the item correct, will be higher for a high ability group than for a low ability group. The requirement of parallel tests, the third problem, is almost impossible to meet in real testing situations. Since the classical test model heavily relies on parallel tests, the violation of the requirement will undermine the accuracy of statistics from classical test theory.

Item response theory (IRT), proposed by Rasch (1980), Birnbaum (1968), Bock (1972), and Lord (1974) among others, offered the possibility of computing invariant statistics. By replacing the measurement model of classical test theory with the estimation model of parameters, IRT made it possible to estimate the person parameter, θ_i, which does not change depending on the item and test parameters, and to estimate invariant item parameters, a_j, b_j and c_j, which do not vary regardless of the level of examinees' ability. The most frequently used model in IRT is the three-parameter logistic (3-PL) model, in which $P_j(\theta)$, the probability of answering the jth item correctly is

$$P_j(\theta) = c_j + \frac{(1-c_j)}{(1+\exp(-1.7a_j(\theta-b_j)))} \qquad (1)$$

where for the jth item, a_j is the slope parameter, b_j is the location or difficulty parameter, c_j is the lower asymptote or guessing parameter, and θ is the ability parameter for a specific examinee. Given the binary response vector, $\mathbf{u}$, the person parameter, θ, can be estimated using the 3-PL model through different ability estimation methods with known or estimated item parameters. Several estimates of θ have been proposed over the past two decades. The MLE(θ) (Birnbaum, 1968), BME(θ) (Samejima, 1969), EAP(θ) (Bock & Aitkin, 1981), and WLE(θ) (Warm, 1989) constitute the major estimation methods. The four major estimators can be classified as either Maximum Likelihood Method (e.g., MLE(θ)) or Bayesian Method (e.g., BME(θ), EAP(θ), and WLE(θ)) (Kim & Nicewander, 1993). EAP(θ) has drawn special attention for its simplicity in computation and accuracy in ability

---

Jwa K. Kim is an Associate Professor of Psychology in the College of Education at Middle Tennessee State University. This work was partly supported by the Faculty Research Grant from Middle Tennessee State University. Please address correspondence regrading the paper to Jwa K. Kim, Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132.

estimation (see Bock & Mislevy, 1982, and Tsutakawa & Soltys, 1988). EAP(θ) was also implemented in the BILOG program (Mislevy & Bock, 1990).

If item parameters are assumed known, the ability of each examinee can be estimated from the likelihood function,

$$L = L(\boldsymbol{u}|\theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \qquad (2)$$

where **u** is the vector of observed item responses; $u_j = 1$ if item j is correctly answered, and 0 if item j is incorrectly answered. Let $P_j = P_j(\theta)$, and $Q_j = 1 - P_j$. If ln(L) is the log likelihood, then

$$\ln(L)' = \sum_{j} \frac{P_j'(u_j - P_j)}{P_j Q_j} \qquad (3)$$

where P_j' and ln(L)' are the partial derivatives of P_j and ln(L) with respect to θ. The maximum likelihood estimate of θ, MLE(θ), can be obtained by setting (3) equal to zero and solving for θ using the Newton-Raphson method or some other suitable numerical strategy.

Unlike MLE(θ), Bayesian Methods assume the prior distribution of ability. If the prior distribution of ability is represented as φ(θ), then the marginal distribution of the item responses is given by

$$L(\boldsymbol{u}) = \int_{-\infty}^{+\infty} L(\boldsymbol{u}|\theta)\phi(\theta)d\theta \qquad (4)$$

where L(**u**|θ) is the conditional likelihood given in (2). The posterior distribution of θ, given **u**, can be derived as

$$p(\theta|\boldsymbol{u}) = \frac{L(\boldsymbol{u}|\theta)\phi(\theta)}{\int_{-\infty}^{+\infty} L(\boldsymbol{u}|\theta)\phi(\theta)d\theta} \qquad (5)$$

using Bayes' Theorem. EAP(θ) estimator which is the mean of the posterior distribution of θ may be expressed as

$$EAP(\theta) = E(\theta|\boldsymbol{u}) = \frac{\int_{-\infty}^{+\infty} L(\boldsymbol{u}|\theta)\phi(\theta)\theta d\theta}{\int_{-\infty}^{+\infty} L(\boldsymbol{u}|\theta)\phi(\theta)d\theta} \qquad (6)$$

The integration in (6) can be approximated using Gauss-Hermite quadrature as

$$EAP(\theta) = E(\theta|\boldsymbol{u}) \doteq \frac{\sum_{k=1}^{q} L(\boldsymbol{u}|X_k)A(X_k)X_k}{\sum_{k=1}^{q} L(\boldsymbol{u}|X_k)A(X_k)} \qquad (7)$$

where X_k and $A(X_k)$ are Gauss-Hermite nodes (abscissas) and their corresponding weights, respectively. These values are available in Gauss-Hermite integration tables (see Stroud & Sechrest, 1966), or they may be approximated by substituting normal deviates for the X_k's and the corresponding, standardized normal densities for the A_k's (the densities are standardized so that they sum to one). For the present study, EAP(θ) will be utilized as the basic model for ability estimation.

According to Equation (1), it is obvious that the probability of answering an item correctly is a function of θ, a_j, b_j, and c_j. The person parameter, θ, is unknown, and it will be estimated through the EAP(θ) method. In the process of ability estimation, item parameters play important roles. Item discriminating parameter, a_j, distinguishes an examinee with high-ability from an examinee with low ability. At any point of ability level, a higher a-value is always desirable. Testing the effect of a-value in ability estimation will be informative.

Item difficulty parameter, b_j, is assumed to have the same distribution as the ability distribution. If other parameters are assumed to remain the same, P_j is the function of b_j. In real testing situations, b_j takes a distribution form instead of a single value, depending on the type of test, for example, a difficult test or an easy test. The present inquiry will investigate the effect of different distributions of b-value.

The guessing parameter, c_j, represents the probability of answering an item correctly by chance alone. Exploring the effect of c-value requires the consideration of other related factors such as the number of alternatives per item, proportionality of alternatives to the number of items, and different ability ranges (Lord, 1980). The effect of c-value, therefore, was excluded from the present study.

Method

Due to the impossibility of closed-form solutions for the conditional means and variances of all θ-estimators, Monte Carlo methods were used to compute these quantities for the θ-estimators (for fixed values of true θ). Furthermore, the fixed-values of true θ were the Gauss-Hermite nodes used in the numerical integration of various functions over the population distribution of θ.

It was assumed that the 3-PL model described the regression of binary item scores on an ability variable that was normally distributed with mean zero and unit standard deviation. The procedure was as follows:

1. Tests of 50 items each, with various values of the *a* and *b* parameters, were specified. Three different values of *a* were chosen (a = .5, 1.0, and 2.0). For each of these, four tests of varying difficulty were specified--a Difficult Test, an Easy Test, a moderately difficult test with a normal-shaped distribution of item difficulties (Normal Test) and a moderately difficult test with a uniform distribution of item difficulties (Uniform Test) (see Table 1 for details). The c-parameters were fixed at .20 for all items assuming five alternatives per item.

Table 1
Number of Items According to Item Difficulty for each 50-Item Test

b-value	Difficult	Easy	Normal	Uniform
-2.0	0	2	0	0
-1.5	0	4	2	7
-1.0	0	8	4	7
-.5	0	16	8	7
0.0	20	20	22	8
.5	16	0	8	7
1.0	8	0	4	7
1.5	4	0	2	7
2.0	2	0	0	0

2. Sixteen quadrature points and weights from a Gauss-Hermite numerical integration table were chosen as true values of θ for each of the 12 test situations. The 16 quadrature points cover 99.99% of the ability distribution, and this number of nodes is generally recognized as sufficient for at least two-place decimal accuracy. The tabled points were multiplied by $(2)^{1/2}$, and the weights divided by $(\pi)^{1/2}$ to scale them correctly for the change in variable needed to integrate a normal distribution (as opposed to the error function integrated by Gauss).

3. For each of the 16 fixed values of θ, 100 response vectors were generated for each test. Each binary element of the response vector, u_j, was generated by drawing a uniformly-distributed random number between zero and one using the TRUE BASIC uniform random function and comparing this number to P_j of the 3-PL model. If a uniformly-distributed random number was smaller than P_j, then u_j was set to one; otherwise, u_j was set to zero. Once completed, this process generated 16 (50x100) binary response matrices for each simulated test.

4. Based on 100 (50x1) response vectors at each θ-value, EAP(θ) was computed as the conditional mean using Equation (7).

5. The bias ($\hat{\theta} - \theta$) was computed along with standard error of the estimation. The bias and standard error beyond the theta value of ±2.8 were truncated due to their small weights.

6. Steps 1 through 5 were repeated twice and the mean of two runs was computed to stabilize random fluctuation. Reported was the mean of the two runs.

Results

Table 2 shows both bias and standard error for each test situation. Standard errors were relatively small and stable. As the a-value increased, the standard error decreased, that is, a highly discriminating item resulted in a more accurate estimation.

A 3x4 (a-value by b-distribution) MANOVA test revealed a significant a-value effect on the combination of bias and standard error, $F(4, 166) = 38.75$, $p = .0001$. Subsequent ANOVAs showed a significant a-value effect on standard error, $F(2, 84) = 94.85$, $p = .0001$, but no significant effect on bias, $F(2, 84) < 1$, $p = .95$. Neither the effect of b-distribution nor the interaction effect were significant.

Although biases were insignificantly different among different a-values and b-distributions, Figures 1 through 3 present very interesting phenomena among the

Table 2
Bias and Standard Error from Different Test

a	θ	Difficult	Easy	Normal	Uniform
.5	-2.8	1.10 (.38)*	.72 (.38)	.93 (.38)	.84 (.40)
	-2.0	.56 (.39)	.33 (.36)	.44 (.37)	.42 (.40)
	-1.2	.19 (.38)	.15 (.39)	.13 (.41)	.17 (.41)
	- .4	.01 (.38)	.01 (.38)	.06 (.39)	.04 (.37)
	.4	- .11 (.37)	- .05 (.39)	- .12 (.34)	- .10 (.41)
	1.2	- .06 (.36)	- .14 (.39)	- .18 (.39)	- .20 (.40)
	2.0	- .32 (.37)	- .39 (.36)	- .37 (.35)	- .37 (.42)
	2.8	- .54 (.32)	- .81 (.33)	- .64 (.36)	- .59 (.36)
1.0	-2.8	1.35 (.28)	.63 (.27)	.87 (.30)	.67 (.27)
	-2.0	.60 (.30)	.19 (.33)	.34 (.32)	.23 (.33)
	-1.2	.01 (.32)	- .02 (.23)	.05 (.33)	- .03 (.32)
	- .4	- .06 (.31)	- .01 (.18)	- .04 (.20)	- .03 (.22)
	.4	- .04 (.19)	.02 (.21)	.00 (.23)	.00 (.25)
	1.2	- .05 (.22)	- .03 (.32)	- .03 (.27)	- .05 (.27)
	2.0	- .05 (.28)	- .30 (.28)	- .17 (.31)	- .09 (.30)
	2.8	- .35 (.27)	- .92 (.17)	- .56 (.22)	- .46 (.26)
2.0	-2.8	1.66 (.24)	.86 (.21)	.95 (.26)	.67 (.15)
	-2.0	.86 (.28)	.31 (.29)	.22 (.30)	- .02 (.24)
	-1.2	.12 (.27)	.07 (.18)	- .08 (.25)	- .01 (.18)
	- .4	- .26 (.24)	- .01 (.07)	- .04 (.12)	.00 (.13)
	.4	.00 (.05)	- .04 (.16)	.00 (.07)	- .01 (.12)
	1.2	.02 (.10)	- .01 (.31)	.06 (.19)	- .02 (.16)
	2.0	.00 (.24)	- .43 (.11)	- .07 (.27)	.00 (.21)
	2.8	- .41 (.16)	- .94 (.00)	- .73 (.11)	- .52 (.08)

* The value in parenthesis is standard error.

different test situations. In general, biases were relatively small for the middle range of the ability distribution, but were large as the theta became extreme values, beyond ±1.5. The second noticeable trend was that as the a-value increased, biases for the middle of the ability distribution became smaller, and biases for the extreme of the distribution became larger. The third phenomenon was not very dramatic but relatively consistent. Except in one instance for the Easy Test when a = 1.0, all tests were more biased at the lower tail of the ability distribution than at the upper tail. Finally, the Difficult Test was always less biased than the Easy Test at the upper tail of the ability distribution, but more biased than the Easy Test at the lower tail of the distribution.

Discussion

This study confirmed a common sense in psychometrics at least partially; better items give better ability estimations. Items with high a-value can be perceived as better items because they distinguish examinees more accurately, which is one of the major tasks of all test items. As the a-value increases, standard error decreases and estimations become more accurate. The reason for the insignificant difference among biases is obvious; the within group variability is larger than the between group variability. From Figures 1-3, one can clearly see that different a-values resulted in different biases depending on the area of the ability distribution. In the middle of the ability distribution, biases become smaller as the a-value increases

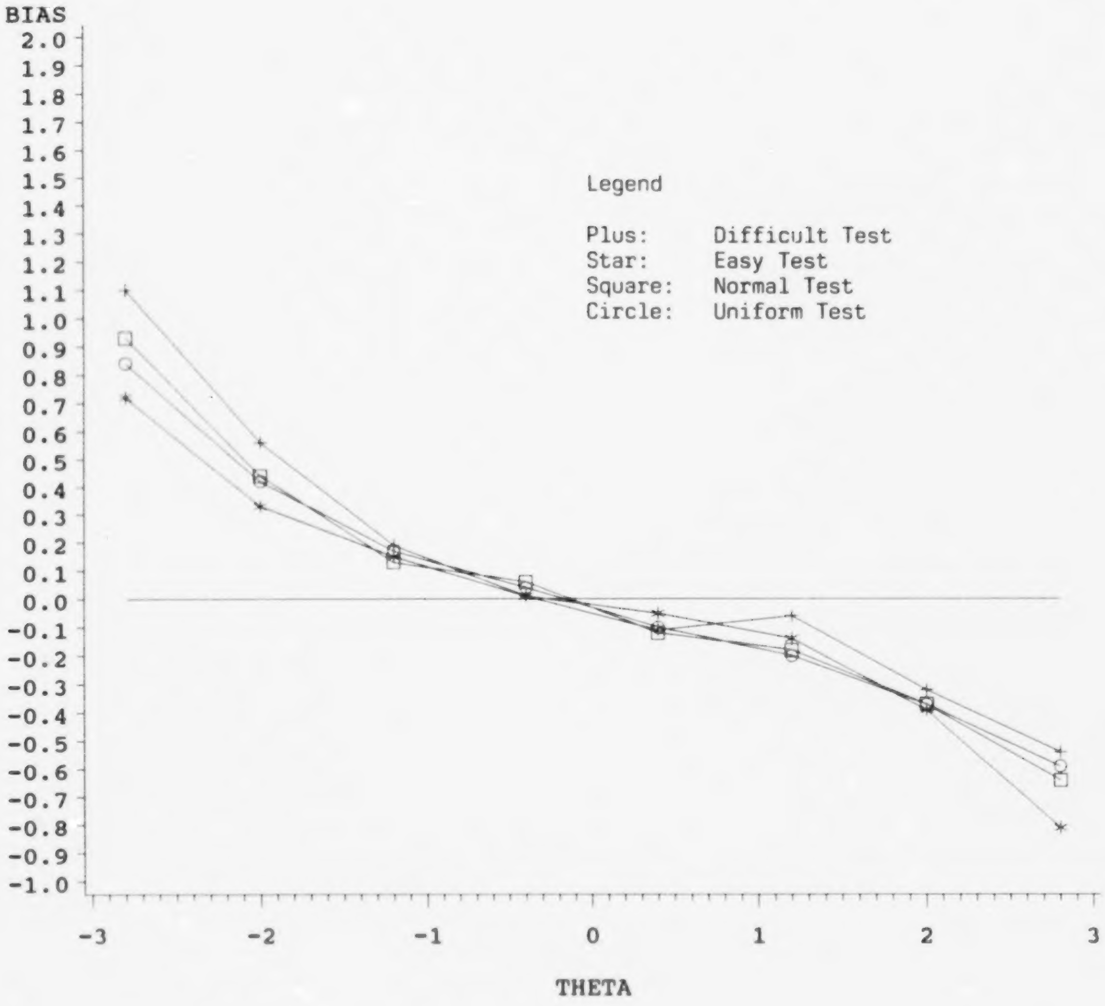


Figure 1. Bias of different tests in ability Estimation, a = .5.

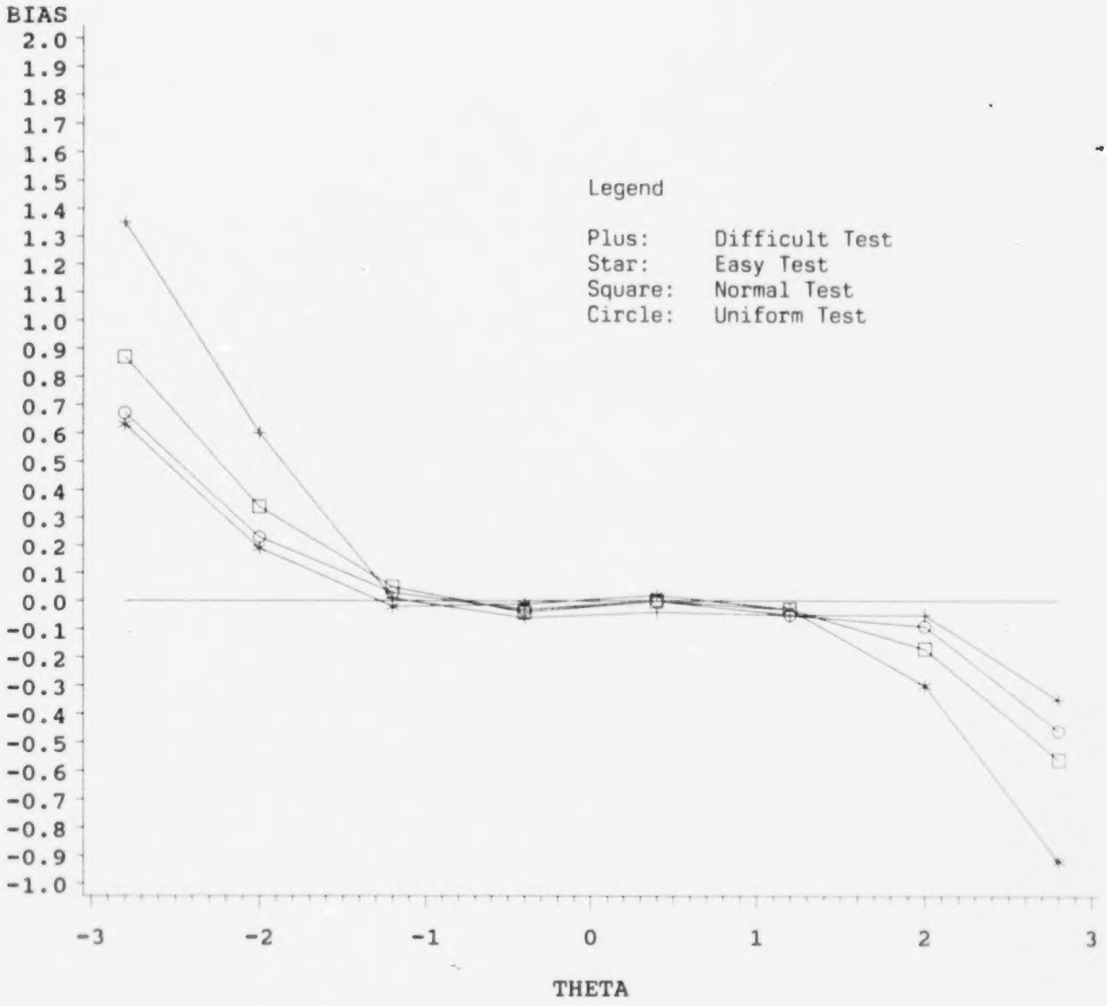


Figure 2. Bias of different tests in ability Estimation, a = 1.0.

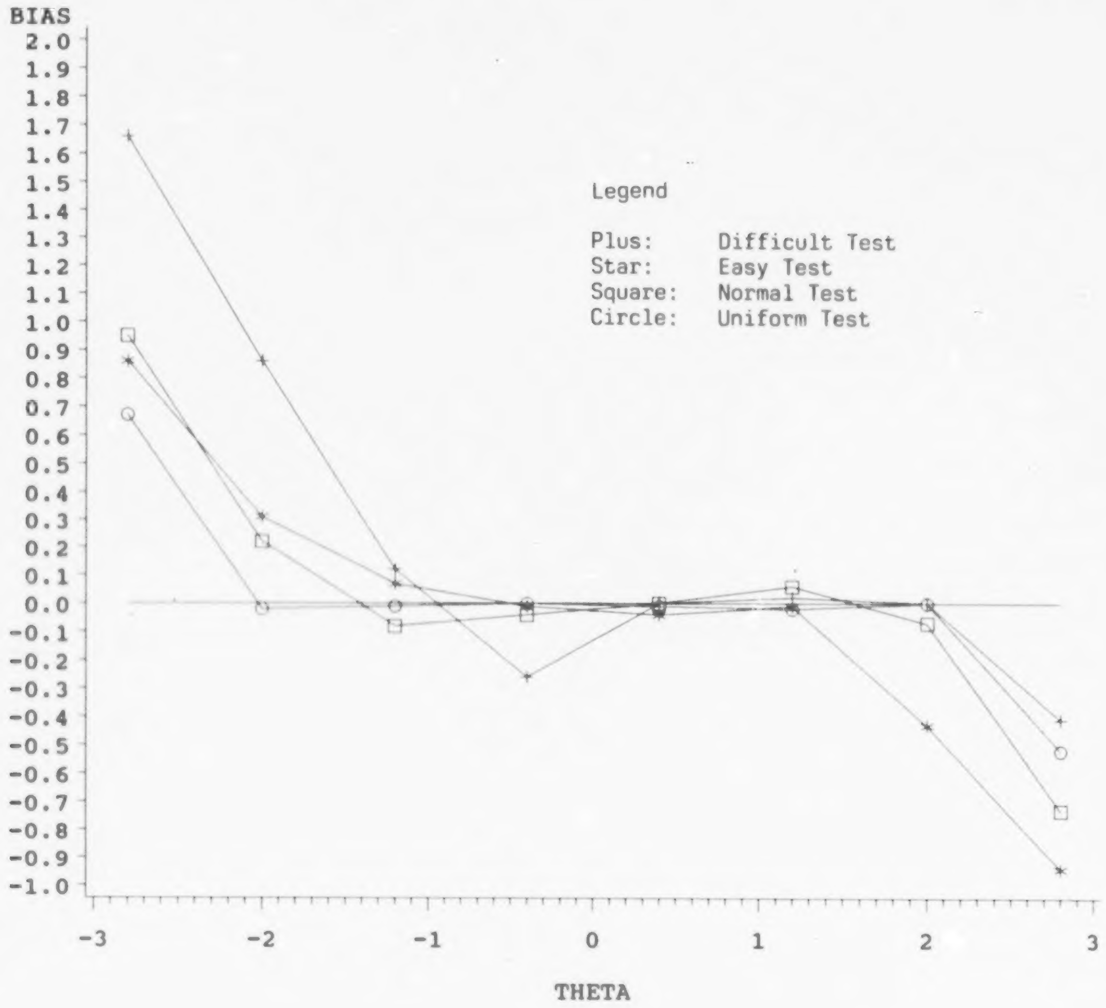


Figure 3. Bias of different tests in ability Estimation, a = 2.0.

from .5 to 2.0. At the extremes of the ability distribution, however, biases become larger with the increase of the a-value.

Although it is statistically insignificant, the Difficult Test is always more biased than the Easy Test at the lower tail of the ability distribution, and less biased than the Easy Test at the upper tail. The appropriate item difficulty, b-value, for the corresponding range of the ability distribution may be the major source of the phenomenon. The Difficult Test has items with the b-values of zero and above, which may give less powerful estimations for the θ-values of zero and below due to the lack of appropriate items for the ability range. The Easy Test which has items with the b-values of zero and below results in larger biases for the upper tail of the ability distribution. This is a very intriguing case which needs further investigation. Due to the same reason, biases for the extremes are always larger than for the middle range of the distribution. A study designed to test the effect of corresponding items for the θ-values would be desirable.

Some practical suggestions can be made for educators, test developers, and test users based on this study. First, use items with high discriminating power if at all possible. These items give a more accurate estimation of ability across all different test situations. Second, use items with difficulty level that corresponds to the examinee's ability level. For the high ability examinees, difficult items result in a better estimation; for the low ability examinees, easy items are better. Third, the Uniform Test performs fairly well in general. A test with about equal numbers of items for all ability levels is recommended for practical test situations.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587-599.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

Mislevy, R., & Bock, R. D. (1990). *PC BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Chicago: Scientific Software.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition). Chicago: University of Chicago Press.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometric Monograph*, No. 17.

Stroud, A. H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice-Hall.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13*, 117-130.

Warm, A. W. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika, 54*, 427-450.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

* * *

MSERA Membership/Renewal Form

(Please print or type)

NAME: ____

TITLE: ____

INSTITUTION: ______________________________________

MAILING ADDRESS: __________________________________
____
____

PHONE: ____________________ FAX: ____________________

ELECTRONIC MAIL ADDRESS:______________ BITNET___ INTERNET___
OTHER___
MSERA MEMBERSHIP: New___ Renewal ___
ARE YOU A MEMBER OF AERA? Yes___ No___
WOULD YOU LIKE INFORMATION ON AERA MEMBERSHIP? Yes___ No___

DUES:	Professional	$15.00	_______
	Student	$10.00	_______

VOLUNTARY TAX DEDUCTIBLE CONTRIBUTION
TO MSER FOUNDATION _______

TOTAL _______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. Dorothy D. Reed (MSERA)
Headquarters, Air University
USAF, 55 LeMay Plaza South
Maxwell AFB, AL 36112-6335

RESEARCH IN THE SCHOOLS
Mid-South Educational Research Association
and The University of Alabama
Office of Research and Service
Post Office Box 870231
Tuscaloosa, AL 35487-0231

TM-21.16

MSERA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and The University of Alabama.

Volume 2, Number 1 **Spring 1995**

Staff Development for Improved Classroom Questioning and Learning 1
J. Jackson Barnette, Jackie A. Walsh, Sandra Orletsky, and Beth D. Sattes

Matching Reading Styles and Reading Instruction 11
Frankie Oglesby and W. Newton Suter

Internalizing/Externalizing Symptomatology in Subtypes of Attention-Deficit Disorder 17
Jose J. Gonzalez and George W. Hynd

Reliability and Validity of Dimensions of Teacher Concern 27
Douglas W. Schipull, Carolyn K. Reeves, and Richard Kazelskis

Preservice Teachers' Views on Standardized Testing Practices 35
Neelam Kher-Durlabhji, Lorna J. Lacina-Gifford, Richard B. Carter, and Randall Jones

Lessons in the Field: Context and the Professional Development of University Participants in an Urban School Placement 41
Janet C. Richards, Joan P. Gipe, and Ramona C. Moore

Locus of Control, Social Interdependence, Academic Preparation, Age, Study Time, and Study Skills of College Students. 55
Craig H. Jones, John R. Slate, and Irmo Marini

The Harrington-O'Shea Career Decision-Making System (CDM) and the Kaufman Adolescent and Adult Intelligence Test (KAIT): Relationship of Interest Scales to Fluid and Crystallized IQs at Ages 12 to 22 Years 63
James E. McLean and Alan S. Kaufman

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of innovative teaching strategies in research/measurement/statistics, descriptions of technology applications in the classroom, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, Office of Research and Service, The University of Alabama, P. O. Box 870231, Tuscaloosa, AL 35487-0231**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1995 by the Mid-South Educational Research Association.

ISSN 1085-5300

RESEARCH IN THE SCHOOLS

Volume 2, Number 1 **Spring 1995**

EDITORS
James E. McLean and Alan S. Kaufman, *The University of Alabama*

PRODUCTION EDITOR
Margaret L. Glowacki, *The University of Alabama*

EDITORIAL ASSISTANT
Anna Williams, *The University of Alabama*

EDITORIAL BOARD
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Michèle Carlier, *University of Reims Champagne Ardenne (France)*
Sheldon B. Clark, *Oak Ridge Institute for Science and Education*
Michael Courtney, *Henry Clay High School (Lexington, KY)*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Kentucky*
Donald F. DeMoulin, *Western Kentucky University*
R. Tony Eichelberger, *University of Pittsburgh*
Daniel Fasko, Jr., *Morehead State University*
Patrick Ferguson, *Arkansas Tech University*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Samuel Hinton, *Eastern Kentucky University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
Randy W. Kamphaus, *University of Georgia*
Jwa K. Kim, Middle *Tennessee State University*
Jimmy D. Lindsey, *Southern University and A & M College*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Peter Melchers, *University of Cologne (Germany)*
Claire Meljac, *Psychologue AU C.H.S. Sainte-Anne (France)*
Soo-Back Moon, *Hyosung Women's University (Korea)*
Arnold J. Moore, *Mississippi State University*
Thomas D. Oakland, *University of Texas*
William W. Purkey, *University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Clemson University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Arkansas State University*
Bruce Thompson, *Texas A & M University*
Anne G. Tishler, *The University of Montevallo*
Wayne J. Urban, *Georgia State University*

GRADUATE STUDENT EDITORIAL BOARD
Vicki Benson, *The University of Alabama*
Ann T. Georgian, *The University of Southern Mississippi*
Jin-Gyu Kim, *The University of Alabama*
Robert T. Marousky, *University of South Alabama*
Jerry G. Mathews, *Mississippi State University*
Dawn Ossont, *Auburn University*
Malenna A. Sumrall, *The University of Alabama*

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 1-10

Staff Development for Improved Classroom Questioning and Learning

J. Jackson Barnette
The University of Alabama

Jackie A. Walsh
Montgomery, Alabama

Sandra R. Orletsky and Beth D. Sattes
Appalachia Educational Laboratory, Inc.

Improving teaching and learning through staff development is an important and much attempted activity. One of the primary tools of the classroom teacher is the use of questioning for the purpose of affecting learning by involving students in meaningful discourse. Improving classroom questioning of teachers by increasing their knowledge and skills is an appropriate focus of staff development. The Appalachia Educational Laboratory, Inc. (AEL) designed and implemented a comprehensive, long-term staff development program, entitled Questioning and Understanding to Improve Learning and Thinking (QUILT), with a goal of increasing and sustaining teacher knowledge of and use of classroom questioning techniques and procedures that produce higher levels of student learning and thinking.

Related Literature

Effective staff development has been associated with certain characteristics. One is that staff development is a process, not an event, that occurs over time (Hord, Rutherford, Huling-Austin, & Hall, 1987). Another important characteristic of effective staff development, as specified by Deal and Kennedy (1983), is the construction of a culture that promotes and sustains change. This includes the development of a shared vision, the use of symbols and metaphors, the development of a common vocabulary, celebration of successes, and other culture-building activities. Fullan (1991) indicates that important variables in successful change efforts are the establishment of a core advocacy group and building of local ownership. Joyce and Showers (1982) indicate that teachers learn and improve performance when provided opportunities to (a) acquire a knowledge base, (b) observe demonstrations, (c) practice new behaviors, and (d) receive feedback on their own performance in the classroom. Staff development is an adult education activity and, as such, Levine (1989) cites as two well-established principles of adult learning the interaction with peers and individual reflection. Lieberman and Miller (1991) suggest that effective staff development is about human development and learning for both students and teachers. They identify "reflective practice" as being critical to successful staff development. McLaughlin (1991) has referred to the need for teaching and learning being co-constructed by teachers and students where there is a more reflective classroom environment in which teachers and students alike have time to think about the content and issues under study.

On the average, teachers dedicate approximately 40% of classroom instructional time to the asking and answering of questions (Doyle, 1986); in machine-gun fashion, they pose an average of 40-50 questions in a typical 50 minute class segment. However, most of these questions are not well prepared and do not serve the purpose of prompting students to think (Dillon, 1988). Ineffective or inappropriate practices include asking questions at only lower cognitive levels (Ornstein, 1987), directing a

---

This work was sponsored wholly or in part by the Office of Educational Research and Improvement, U.S. Department of Education. Its contents do not necessarily reflect the views of OERI, the Department, or any other agency of the U.S. Government. J. Jackson Barnette is professor of educational research at The University of Alabama, Tuscaloosa, Alabama. Jackie A. Walsh is an independent educational consultant, Montgomery, Alabama. Sandra R. Orletsky is director of the school governance and administration program of the Appalachia Educational Laboratory, Inc., Charleston, West Virginia. Beth D. Sattes is senior research and development specialist of the Appalachia Educational Laboratory, Inc., Charleston, West Virginia. Correspondence regarding this paper should be addressed to Dr. J. Jackson Barnette, Area of Professional Studies, The University of Alabama, P. O. Box 870231, Tuscaloosa, AL 35487-0231.

disproportionate percentage of questions toward a limited number of students (Jones, 1990), or waiting too little time (Wait Time I) after asking a question or before reacting to the student's response (Wait Time II), typically one second or less (Rowe, 1986).

QUILT Overview

QUILT is a staff development program for classroom teachers which is designed to incorporate many of the components for effective staff development described in the literature. It attempts to help teachers improve the quality of the questions they pose and increase the use of behaviors that facilitate involvement and learning in the process of classroom discourse.

QUILT has four components: induction training, collegiums, partnering, and independent study and analysis. Induction training is an 18-hour program, conducted by trainers trained by AEL, where participants are provided research-based knowledge and theory, as well as frequent opportunities to practice effective questioning techniques. Data from induction training evaluation indicate that the 18-hour induction leads to acquisition of a knowledge-base, a common vocabulary, and a culture which produces a degree of bonding among participants that does not occur during shorter sessions. The selection of QUILT as the program acronym had a major affect on development of participant bonding. Not only were the program components and training designed around the development of a quilt, but the sharing of stories about family quilts by participants in informal, getting acquainted sessions was a major factor in development of culture and program ownership.

During the school year, teachers and administrators meet seven times in forums, referred to as collegiums, designed to review information about questioning and reinforce changes in teacher questioning behaviors. Each focuses on greater understanding and reinforcing of particular questioning skills and behaviors. Partnering involves teams of peer teachers in ongoing, mutual support activities within the school. These activities include visiting each other's classrooms to observe and monitor progress in questioning and to provide support and encouragement. Throughout the year participants read independently, practice their skills, and compile data on their own classroom behaviors and student responses.

QUILT differs in significant ways from the approaches to staff development frequently employed by schools. First, QUILT treats staff development as a long-term commitment. Research indicates that only a small percentage of teachers, perhaps as low as 10%, change their behavior in response to a training program unless lectures or seminars are reinforced by feedback in a classroom setting (Joyce & Showers, 1982). QUILT is a multi-year program and the "partnering" or peer coaching approach is central to its design.

Second, QUILT represents a "whole-school" approach to staff development. Because questioning is a generic educational activity, improving questioning skills is relevant across disciplines from kindergarten to 12th grade. The partnering approach reflects this generic quality since teachers across subject areas can work together to improve questioning skills. Third, QUILT is student-centered. While it is fashionable to make this claim for almost any program, the entire purpose of the QUILT five-stage model is to stimulate student thinking, particularly higher order thinking. Stage 1 relates to preparing the question. It includes identifying the instructional purpose, determining the content focus, selecting the appropriate cognitive level, and considering wording and context. Stage 2 relates to presenting the question. It includes indicating the response format, asking the question, and respondent selection. Stage 3 relates to prompting student response. It includes pausing after asking the question, assisting nonrespondents, and pausing after student response. Stage 4 relates to processing the student response. It includes provision of appropriate feedback, expanding and using correct responses, and eliciting student reactions and questions. Stage 5 relates to critiquing the questioning episode, including analyzing the question, mapping respondent selection, evaluating student response patterns, and examining teacher and student reactions.

While there are several factors cited in the literature associated with effective staff development, there are few examples of empirical studies which examine effects of varying the presence or absence of these factors. During the field test year, three levels of QUILT implementation were designed with varying degrees of presence of literature-based staff development factors. It was hypothesized that groups with higher levels of presence of these staff development effectiveness factors would have higher levels of knowledge gain and higher use of actual classroom behaviors associated with effective questioning practices.

Methodology

Because QUILT was whole-school based, it was not possible to assign individual teachers to one of the three conditions. However, schools were randomly assigned into one of the three conditions. Schools at different levels (elementary, middle, and secondary schools) were represented in each condition. Condition A schools completed the full QUILT program which included 18 hours of induction training, collegiums, partnering, and other independent study activities. Condition B schools completed only the 18-hour induction training, but had no systematic year-long continuation activities. Condition C schools received only a 3-hour orientation session related to QUILT questioning concepts, the typical staff development mode. Forty-one schools from 13 districts in Kentucky, Tennessee, Virginia, and West Virginia participated in the field test.

An extensive research design was developed to assess QUILT effectiveness (Barnette & Sattes, 1991). This included pre- and post-QUILT assessment of teacher knowledge, teacher attitudes, student attitudes, and classroom questioning practices. In addition, evaluation of all aspects of program delivery and implementation was conducted. The aspects of QUILT research reported here are related to the effects of QUILT on teacher knowledge of effective classroom questioning practices, based on a paper-pencil test, and actual classroom questioning practices, based on the videotaped observation and coding of a sample of QUILT teachers.

Since schools rather than teachers were randomly assigned to treatment conditions and since teachers were randomly selected from these groups for the videotape sample, it was important to compare the teachers in the three conditions relative to certain demographic characteristics to ensure comparability. Variables compared included type of position, grade level, subject taught, gender, ethnicity, age, highest academic degree, years of experience, and types of staff development attended previously. There were no pre-QUILT significant differences on these variables between the groups. Relative to percentage of teachers as compared with other professionals, in the three conditions there were 91% in Condition A, and 89% in each of the other two conditions. The only variable where there was a discrepancy was for grade level. Condition A had a lower percentage of middle school teachers (9%) and a higher percentage of secondary teachers (47%) as compared with the other two conditions; there were 25% middle school teachers in Condition B and 21% in Condition C, and 29% secondary teachers in Condition B and 41% in Condition C. This was not surprising since schools were assigned to conditions, and school sizes vary. When examining the subjects taught, there was representation of more than 20 different subjects in each condition. The conditions had male and female distributions of 26% male and 74% female in Condition A, 22% male and 78% female in Condition B, and 25% male and 75% female in Condition C. Most of the participants were Caucasian (94% in Condition A, 95% in Condition B, and 96% in Condition C). The balance by age category, by highest degree, by years of experience, and types of previous staff development was highly consistent across the three conditions. In addition, pretest means were compared between conditions on each of the QUILT outcome variables using ANOVA, and no significant preQUILT differences were found.

Assessing Knowledge of Classroom Questioning

The Questionnaire on Effective Classroom Questioning (QECQ) measures teacher knowledge about and understanding of classroom questioning and its relationship to student learning and thinking. QUILT staff and consultants developed this instrument after an extensive search failed to turn up an existing instrument. Criteria for development of this instrument included correspondence between content of test items and QUILT objectives, a sufficient degree of difficulty to yield a reasonable level of score variability needed to assess pre to post knowledge change, and an acceptable level of reliability.

An extensive review of current research on classroom questioning was the basis for the items on the QECQ. Items were identified for six subscales of effective questioning (general concepts), teacher feedback and reaction, discussion vs. recitation, respondent selection and response formats, cognitive levels, and wait times. Two versions of the QECQ were field tested and revised based on responses of teachers similar to those who participated in QUILT. The final version, which was used in the field test of QUILT, was a 49-item, multiple-choice instrument. Internal consistency reliability for this instrument was .76, based on the posttest scores of more than 1,200 QUILT participants. The instrument has a high level of difficulty. The guess score was 28.6%. Pretest scores for all QUILT respondents had a mean of 46.3% correct. Only teachers with scores on the QECQ at both pre and post times are included in this analysis ($n = 789$).

Assessing Classroom Questioning Behaviors

The Classroom Questioning Observation Instrument (CQOI) was developed for the purpose of collecting data on teachers' classroom questioning behaviors. More specifically, the behaviors of interest included number of teacher-initiated questions, use of Wait Time I (the time a teacher waits to acknowledge a student's response), use of Wait Time II (the amount of time a teacher waits before reacting to a student's answer), cognitive levels (recall, utilization, or creation) of questions and student answers, manner of designating students to answer questions, and use of various types of desirable and undesirable teacher responses or reactions (Barnette, Sullivan, Orletsky, & Sattes, 1993).

Because participating teachers were spread out over four states and in an attempt to reduce obtrusiveness of an actual observer in the classroom, it was decided to have 15-minute videotapes recorded, which would be reviewed and coded by trained coders. Also, it was not feasible, for both cost and time limitations, to videotape and code all participating teachers in the four states. A randomly selected sample of QUILT teachers was identified. The sample size was based on two factors, having a reasonable level of statistical power and being practical in terms of costs of collecting and coding videotapes. According the Hinkle, Wiersma, and Jurs (1994), a sample size of 32 per group has the power of .95 in detecting a significant difference, at an effect size of one, in a three-group situation where alpha is set at .05. Thus, a minimum of 96 teachers was needed in the sample. Realizing that there would be some attrition at either the pre or post times, it was decided to sample 135 teachers, which was slightly higher than 10% of the total participant group. Districts were then asked to videotape the selected teachers two times, once in the spring before QUILT training and again in the spring at the end of the QUILT field test year. Only teachers with both pre and post videotapes are included in this analysis ($n = 95$).

Dr. Debra Sullivan, the CQOI developer, used prior knowledge of other classroom observation instruments, QUILT materials, and classroom visits to design the instrument. Throughout the instrument's formative stages of development, the developer visited classrooms and collected data using draft versions of the instrument. Using this process, not only was it possible to assess specific research questions, but also "real life" usability in classroom situations was assured. Meetings were held with AEL staff and consultants to ensure a match between the research design and the teacher behavior data collection device, increasing the level of content validity.

For logistic reasons, it was decided to have all coders living in the Charleston, WV area. Four teachers were selected by the CQOI developer to participate in coder training. All of the selected teachers were considered extremely capable and competent teachers who represented several major curriculum areas including language arts, social studies, mathematics, science, and foreign language. Coders were trained using a variety of methods, including group sessions as well as independent work. During the training sessions, coders were acquainted with the QUILT program and its research design; were familiarized with the CQOI in terms of format, definitions, and manner of completion; practiced coding transcripts of classroom sequences featuring questioning interactions between teachers and students; and practiced coding videotapes of classroom episodes. During the coder training, CQOI codes and their definitions were discussed and defined more clearly, thus increasing levels of coder validity and agreement.

Since 15-minute videotapes of classroom teaching episodes were used rather than direct observation, coder speed was not an area of concern. Coders were able to replay the tape to check coding for accuracy and reconsideration. Therefore, only accuracy in coding classroom questioning behaviors was necessary to determine coder level of agreement. The extent of consistency was established by comparing coder responses with those of the CQOI developer on the same videotape. Agreement of coding ranged from 90 to 94%, with an average agreement of 92%. Coders did not know the teachers who were observed, nor did they know which QUILT condition the teacher represented.

Each questioning episode is recorded in terms of whether it was teacher or student initiated. For teacher initiated questions, whether the teacher designated a student to answer before or after asking the question is then recorded. The level of question is recorded as being recall, utilization, or creation. Wait Time I, the time a teacher waits before acknowledging a student response to an initial question, is recorded by checking the number of seconds. The student answering, whether the one designated before or after the question was asked, is recorded. The number of students responding is recorded as one, more than one, or whole class (choral response). The level of student answer is recorded as being recall, utilization, or creation. The student answer is also recorded as being correct, partially correct, wrong, no answer, inappropriate response, and whether the student asks for clarification or extends his/her answer.

Wait Time II, the time the teacher waits before reacting to the student's answer, is recorded in seconds. The teacher reaction is recorded as being positive feedback, praise, negative feedback, corrective feedback, criticism, or no feedback given. In addition, other teacher behaviors are recorded including whether the teacher probes, repeats or rephrases the question, repeats or rephrases the student answer, uses the student response in discussion or new questions, and/or redirects the question.

Data Analysis

Three different groups were included in this analysis: Condition A (full QUILT model including induction and collegiums), Condition B (QUILT induction without collegiums), and Condition C (QUILT 3-hour awareness session only). Data were analyzed using several SAS procedures (SAS Institute, Inc.; 1989a, 1989b) (SAS is a registered trademark of SAS Institute Inc., Cary, NC). For each QUILT variable, the following analyses were conducted:

1. Univariate summary statistics were computed for pretest results, posttest results, and post-pre test differences. Included were tests for normality and provision of data for computation of F_{max} statistics for checking analysis of variance assumptions. These results were also used to compute effect sizes. The pretest standard deviation for all participant scores in the three conditions was used as the base for the effect size. The posttest minus pretest means were divided by the overall pretest standard deviation to obtain the effect size.

2. The GLM (general linear model) procedure was conducted as a mixed design, with a between subjects factor (condition) and a within subjects factor (testing time). Of primary concern were three planned follow-ups. Since these comparisons were in the planned mode, significant main effect or interaction of condition and time were not required to conduct these follow-ups.

3. The first follow-up procedure involved the comparison of pre- and posttest means within each condition. These were compared using directional, dependent *t* tests with alpha set at .05 and adjusted with a Bonferroni correction.

4. The second follow-up procedure involved the comparison of posttest means of Condition A with each of the other groups (A with B and A with C). These were compared using directional, Dunnett *t* tests with alpha set at .05. In this case, Condition A was compared with each of the other groups.

5. The third follow-up procedure involved the comparison of the pre- to posttest change mean of Condition A with each of the other groups (A with B and A with C). These were compared using directional Dunnett *t* tests with alpha set at .05.

Tests of the normality and homogeneity of variance assumptions on the QECQ indicated no significant departures from normality. There were some significant differences in group variances. These, however, were largely the function of large sample sizes. There was a significant departure from normality for the Wait Time II observation variable. There were no significant homogeneity of variance differences for the observation data. Since there is not a satisfactory nonparametric alternative to the mixed ANOVA design and follow-ups and since ANOVA is robust to violations of these assumptions, the two-way ANOVA procedure was used.

Results

Teacher Knowledge of Effective Classroom Questioning

Table 1 presents results on the QECQ. All three conditions had significant pre to post increases on the total QECQ score. Condition A had a significantly higher mean than both of the other conditions at posttest. Also, Condition A demonstrated significantly higher pre to post changes than both of the other conditions. The pre to post effect size was +1.17 for Condition A, +0.64 for Condition B, and +0.24 for Condition C.

Clearly, the largest pre to post change was observed on the wait time subscale for all three conditions, an effect size of +1.30 for Condition A, +.89 for Condition B, and +.46 for Condition C. Other subscales with higher than .5 effect sizes for Condition A were cognitive levels (+.73) and characteristics of effective questions (+.61). Condition A had significant pre to post changes on all six subscales, Condition B had significant pre to post changes on five of the subscales, and Condition C had significant pre to post changes on only one of the subscales. At posttesting, Condition A had higher subscale means than Condition B on three of the subscales, and Condition A had higher subscale means on all six of the subscales when compared with Condition C. In addition, Condition A had significantly higher pre to post changes than Condition B on three of the subscales and significantly higher pre to post changes on all six subscales as compared with Condition C.

Classroom Questioning Behaviors

Table 2 presents results for the selected classroom questioning behaviors based on the coding of the videotapes. During the 15-minute videotape, the number of teacher initiated questions was recorded. The desirable change was that there be a decrease on this variable. QUILT stressed the need to have fewer, better planned questions. All three groups had reductions in the number of teacher questions. This reduction was significant for Condition A, with an effect size of -.65. At post, Condition A had a significantly lower number of teacher initiated questions than condition B. There were no significant differences between Condition A and Conditions B or C relative to the degree of change between pre and posttest number of questions asked.

Table 1
Pre and Post Comparisons on Questionnaire About Effective Classroom Questioning.
Percent Correct by Subscale and Total

		Treatment Condition						
		A, n = 297		B, n = 200		C, n = 292		Group Differences in Means* Post Change
Subscale		Pre	Post	Pre	Post	Pre	Post	
Effective	*M*	40.5	52.9	42.3	46.2	40.8	41.1	A>B A>B
questioning	*SD*	20.4	22.1	20.2	21.6	19.9	19.9	A>C A>C
SD_{pre}= 20.2	*ES*	.61		.19		.01		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Feedback	*M*	49.3	54.5	49.1	52.5	49.5	48.4	A>C A>C
and reaction	*SD*	13.6	15.0	12.7	14.8	13.0	13.9	
SD_{pre}= 13.1	*ES*	.39		.25		-.08		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Discussion vs.	*M*	37.1	46.1	37.6	42.6	35.8	39.1	A>C A>C
recitation	*SD*	20.9	20.8	22.1	21.2	20.7	20.5	
SD_{pre}= 21.2	*ES*	.43		.24		.16		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Selection	*M*	44.7	50.9	45.1	47.9	40.7	43.1	A>C A>C
and format	*SD*	17.2	17.0	18.1	18.7	16.7	17.2	
SD_{pre}= 17.3	*ES*	.36		.16		.14		
Post-Pre Diff.*		Post>Pre		nsd		nsd		
Cognitive	*M*	50.7	64.1	52.3	57.8	48.5	51.1	A>B A>B
levels	*SD*	18.6	20.1	18.2	19.5	17.7	19.3	A>C A>C
SD_{pre}= 18.2	*ES*	.73		.30		.15		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Wait time	*M*	50.5	78.8	49.4	68.9	46.3	56.2	A>B A>B
	SD	22.7	21.5	22.1	24.0	20.5	23.4	A>C A>C
SD_{pre}= 21.8	*ES*	1.30		.89		.46		
Post-Pre Diff.*		Post>Pre		Post>Pre		Post>Pre		
Total QECQ	*M*	46.8	58.2	47.2	53.4	45.1	47.4	A>B A>B
	SD	10.3	12.3	9.7	12.7	9.1	10.5	A>C A>C
SD_{pre}= 9.7	*ES*	1.17		.64		.24		
Post-Pre Diff.*		Post>Pre		Post>Pre		Post>Pre		

*p <.05 after applying Bonferroni correction

Table 2
Pre and Post Comparisons on QUILT Observation Variables

Variable		A, n = 37 Pre	A, n = 37 Post	Treatment Condition B, n = 28 Pre	Treatment Condition B, n = 28 Post	C, n = 30 Pre	C, n = 30 Post	Group Differences in Means* Post Change
Number, tchr.	*M*	41.4	31.0	44.9	40.5	43.3	36.3	A<B
questions[1]	*SD*	15.8	14.5	17.4	13.8	15.5	14.4	
SD_{pre}= 16.1	*ES*	-.65		-.27		-.44		
Post-Pre Diff.*		Post<Pre		nsd		nsd		
Mean wait	*M*	0.90	1.70	0.83	1.32	0.74	0.80	A>C A>C
time I	*SD*	0.58	1.47	0.53	1.03	0.74	0.93	
SD_{pre}= .62	*ES*	1.29		.79		.11		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Wait time I,	*M*	12.8	25.0	11.1	20.7	10.1	11.5	A>C A>C
3 sec./more, %	*SD*	11.9	24.9	10.1	19.5	14.8	16.5	
SD_{pre}= 12.3	*ES*	.99		.78		.11		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Mean wait	*M*	.08	.43	.03	.18	.06	.16	A>B A>C
time II	*SD*	.12	.53	.04	.31	.14	.33	A>C
SD_{pre}= .11	*ES*	3.13		1.37		.92		
Post-Pre Diff.*		Post>Pre		Post>Pre		nsd		
Wait time II,	*M*	.52	2.98	.10	.59	.59	.97	A>B A>C
3 sec./more, %	*SD*	1.28	6.73	.51	1.61	2.06	4.57	A>C
SD_{pre}= 1.44	*ES*	1.72		.34		.26		
Post-Pre Diff.*		Post>Pre		nsd		nsd		
Quest. above	*M*	31.0	41.3	41.0	39.2	26.3	32.0	
recall, %	*SD*	23.3	27.8	24.8	30.1	22.0	22.7	
SD_{pre}= 23.8	*ES*	.43		-.07		.24		
Post-Pre Diff.*		Post>Pre		nsd		nsd		
Answer above	*M*	28.4	37.6	38.2	35.9	25.2	29.6	
recall, %	*SD*	21.7	25.4	23.3	28.5	21.8	20.4	
SD_{pre}= 22.6	*ES*	.41		-.10		.19		
Post-Pre Diff.*		Post>Pre		nsd		nsd		
Quest. redir.	*M*	14.1	23.2	20.6	19.4	18.1	12.3	A>C A>B
to other(s), %	*SD*	14.5	19.9	16.7	14.9	15.0	14.5	A>C
SD_{pre}= 15.4	*ES*	.59		-.08		-.37		
Post-Pre Diff.*		Post>Pre		nsd		nsd		
Student desig.	*M*	84.1	90.8	83.1	85.3	83.5	89.4	A>B
aft. quest., %	*SD*	12.8	9.3	23.2	11.2	14.4	11.2	
SD_{pre}= 16.8	*ES*	.40		.13		.35		
Post-Pre Diff.*		Post>Pre		nsd		Post>Pre		
Tchr. repeats	*M*	62.4	54.6	60.5	55.9	59.4	61.5	
answer, %[1]	*SD*	18.9	28.5	14.3	17.9	20.9	25.5	
SD_{pre}= 18.2	*ES*	-.43		-.25		.11		
Post-Pre Diff.*		Post<Pre		nsd		nsd		

[1] Predicted, desired direction for these variables is negative.
*p < .05 after applying Bonferroni correction

BEST COPY AVAILABLE

Wait Time I was compared among the conditions. It is recommended that Wait Time I be three seconds or longer. It was predicted that this variable would increase. Two variables were observed, the mean Wait Time I and the percent of time the teacher waited three seconds or longer. Both Conditions A and B had significant increases in this variable, with effect sizes of +1.29 for Condition A and +.79 for Condition B on the mean Wait Time I and effect sizes of +.99 for Condition A and +.78 for condition B on the percent of time Wait Time I was three seconds or longer. Condition A had a higher mean at post, as well as significantly higher pre to post change as compared with Condition C.

It was predicted that Wait Time II would increase. It is recommended that this time be three seconds or longer. While the level of the use of Wait Time II is very low, Conditions A and B had significant pre to post mean increases, with effect sizes of +3.13 for Condition A and +1.37 for Condition B. Only Condition A had a significant increase in the percent of time teachers used Wait Time II at three seconds or longer, with an effect size of +1.72. At post, Condition A was significantly higher than both Conditions B and C on the mean Wait Time II and the percent of time Wait Time II was three seconds or higher. Condition A had a significantly higher pre to post change as compared with Condition C on both of these variables. While there were significant differences observed, the use of Wait Time II was still much lower (.43 seconds) than recommended (3 seconds).

The percent of time teacher initiated questions were above recall cognitive level was determined. An objective of QUILT training was to increase the frequency of higher level questions. Condition A was the only group to have a significant pre to post change, with an effect size of +.43. There were no significant differences between the groups at post nor relative to pre to post changes.

The percent of time the student's answer was above recall cognitive level was determined. An objective of QUILT training was to increase the frequency of higher level answers. Condition A was the only group to have a significant pre to post change, with an effect size of +.41. There were no significant differences between the groups at post nor relative to pre to post changes.

The percent of time a teacher redirected a question to other student(s) was determined. A QUILT objective was for this to increase. Condition A had a significant pre to post change with an effect size of +.59. At post, Condition A had a significantly higher percent than Condition C, and Condition A had significantly higher pre to post change than both Conditions B and C.

The percent of time the teacher designated the student to answer a question after it was asked was determined. It was a QUILT objective to increase this practice because, often when the student is designated prior to the question rather than after the question, other students, since they feel they are not involved, reduce or discontinue their involvement in the interaction; they are "off the hook." Both Conditions A and C had significant pre to post changes, with the effect size for Condition A at +.40 and for Condition C at +.35. At post, the Condition A mean was significantly higher than Condition B.

Another variable, which QUILT was designed to decrease was the percent of time a teacher repeats the student answer. Often when this happens other students take this as acknowledging the response as being correct and then there is no need to continue thinking. If the teacher "has" the answer, students often tune-out. Condition A was the only one to have a significant pre to post reduction in this behavior, with an effect size of -.43.

Summary of Differences

On the QECQ, there were significant pre to post differences on the total score and all subscales for Condition A, on the total score and five of the subscales for Condition B, and on the total and one of the subscales for Condition C. On the 10 observation variables, there were significant pre to post changes on every one of the variables for Condition A, Condition B had significant differences on three of the variables, and Condition C had significant differences on one of the variables.

On the QECQ, Condition A had significantly higher pre to post change on the total score and all six subscales compared with Condition C and higher pre to post change on the total and three of the subscales as compared with Condition B. On the observation variables, Condition A had significantly higher predicted pre to post change than Condition C on the variables of mean and percentage at 3 seconds or more on both Wait Time I and Wait Time II, and question redirection to other student(s). Condition A had significantly higher predicted pre to post change than Condition B on the variable of question redirection to other student(s).

Threats to Validity

Conducting research in field settings has high potential for threats to four types of validity (Cook & Campbell, 1979), statistical conclusion, internal, construct, and external. While it is not possible to discuss each of the more than 30 recognized threats to validity and how each was controlled or minimized, conscious efforts were made to account for many of these threats. Rival hypotheses were controlled, or effects minimized, by using powerful statistical methods with planned, directional hypotheses; random assignment of schools to treatment conditions; random selection of subjects for observational data collection; within-school singular treatment rather than multiple treatments occurring within the same school setting; well designed materials and high quality training of district-based trainers; and replication of treatments in several settings.

Conclusion and Implications

The effectiveness of QUILT as an example of focused and integrated staff development was tested in a large scale experiment based on comparison of groups randomly assigned to one of three conditions and collection of data with respectable levels of validity and reliability. Based on these results, effective staff development can be accomplished if characteristics identified in the research literature are present. Condition A was designed to specifically incorporate aspects of effective staff development including using a long-term change process rather than a single event; culture-building which promotes and sustains change; collegial interaction and support; and a process which provides opportunities for teachers to acquire a knowledge base, observe demonstrations, practice new behaviors, and receive feedback on their own performance in the classroom. Clearly, the teachers who participated in Condition A, the full QUILT implementation, had greater increases in knowledge of classroom questioning and more positive classroom behaviors than teachers who participated in less than the full QUILT implementation. Several important variables associated with classroom questioning may be influenced in a positive way by focused staff development, and improving teaching practices should have a concurrent improvement in student classroom involvement, thinking, and learning.

References

Barnette, J. J., & Sattes, B. D. (1991). *QUILT data collection coordinator's manual*. Charleston, WV: Appalachia Educational Laboratory.

Barnette, J. J., Sullivan, D., Orletsky, S. R., & Sattes, B. D. (1993, April). Observation of teacher classroom questioning behaviors. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Cook, D. T., & Campbell, D. T. (1979). *Quasi-experimentation: Design analysis issues for field settings*. Chicago: Rand-McNally.

Deal, T. E., & Kennedy, A. A. (1983). Culture and school performance. *Educational Leadership, 40*(5), 14-15.

Dillon, J. T. (1988). *Questioning and teaching: A manual of practice*. New York: Teachers College Press.

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Research on teaching* (3rd ed.) (pp. 392-431). New York: Macmillan.

Fullan, M. G. (1991). *The new meaning of educational change*. New York: Teachers College Press.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). *Applied statistics for the behavioral sciences* (3rd ed.). Boston: Houghton Mifflin.

Hord, S. M., Rutherford, W. L., Huling-Austin, L., & Hall, G. E. (1987). *Taking charge of change*. Alexandria, VA: Association for Supervision and Curriculum Development.

Jones, M. G. (1990). T-zone, target students and science classroom interactions. *Journal for Research in Science Teaching, 27*(7), 631-660.

Joyce, R. B., & Showers, B. (1982). The coaching of teaching. *Educational Leadership, 40*(1), 4-10.

Levine, S. L. (1989). *Promoting adult growth in schools: The promise of professional development*. Boston: Allyn & Bacon.

Lieberman, A., & Miller, L. (1991). Revisiting the social realities of teaching. In A. Lieberman & L. Miller (Eds.). *Staff development for education in the 90's* (pp. 92-112). New York: Teachers College Press.

McLaughlin, M. W. (1991). Enabling staff development: What have we learned? In A. Lieberman & L. Miller (Eds.), *Staff development for education in the 90's* (pp. 61-82). New York: Teachers College Press.

Ornstein, A. C. (1987). Questioning: The essence of good teaching. *NASSP Bulletin, 71*(499), 71-79.

Rowe, M. B. (1986). Wait time: Slowing down may be a way of speeding up. *Journal of Teacher Education, 37*(1), 43-48.

SAS Institute, Inc. (1989a). *SAS/STAT user's guide: Volume 1, ANOVA-FREQ, Version 6, 4th ed.* Cary, NC: The Author.

SAS Institute, Inc. (1989b). *SAS/STAT user's guide: Volume 2, GLM-VARCOMP, Version 6, 4th ed.* Cary, NC: The Author.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 11-15

Matching Reading Styles and Reading Instruction

Frankie Oglesby
Judson College

W. Newton Suter
University of Arkansas at Little Rock

This research tested the hypothesis that reading instruction designed to incorporate students' preferred reading styles resulted in greater achievement gain when compared to comparable controls. One hundred ninety-eight 3rd and 6th graders were pretested using a standardized reading achievement test. Half of the participants were administered the Reading Style Inventory (Carbo, 1982) and were provided with reading instruction using methods, strategies, and materials that matched their preferred style. Six months later, all students were posttested using the same reading test. Results supported the hypothesis that incorporating reading styles in reading instruction resulted in greater achievement gain.

Today's classroom teachers attempt to educate more children with varying levels of ability and diversified cultural backgrounds than ever before. School children are being exposed for the first time to highly stimulating technology by increasingly reflective teachers who are benefitting from the past two decades of productive research on individual differences and instructional methods that focus on classroom processes (aptitude-treatment interactions, or ATIs). To bring today's students into a confining environment and group them in an educationally sensible way is virtually impossible unless we assess the individual in order to identify exactly how he or she is likely to learn most effectively (Bennett, 1979). Koons' (1977) reminder that "schools should be made to fit the student, not students made to fit the schools" (p. 701) is more relevant than ever before.

One promising approach to "fitting" students involves accommodating their individual learning styles. Learning style research began in the early 1970s and is now receiving widespread attention. The Carbo (1988a) and Dunn (1988) reviews revealed that increasing numbers of studies are showing that instruction which accommodates learning styles yields higher achievement and improved attitudes across grade levels and content areas.

Research in learning styles and reading achievement acknowledges that the identification of a student's learning style may be paramount in eliminating reading failure. A promising methodology for improving reading achievement is the use of reading style diagnosis and prescription (Carbo, 1990). Materials, methods, and instruction that do not match a student's reading style, strengths, and preferences dramatically increase the severity and pervasiveness of reading problems (Carbo, 1984). As teachers become more competent in diagnosing individual learning styles and providing instructional strategies to capitalize on them, it is probable that children increasingly will be grouped according to both performance levels and to methods that maximize their reading achievement (Carbo, 1980). The task of the educator is to determine how the strengths of the individual's learning style can be utilized and the influence of weaknesses reduced.

Although few would argue about the instructional value of capitalizing on students' strengths, there is disagreement regarding evidence to support matching reading styles and instructional methods (O'Neil, 1990; Stahl, 1988). Although Stahl (1988) noted that it may be impossible to conduct a "flawless" study, his review of research matching reading style with reading instruction questioned whether the research base was "sufficiently rigorous or valid" to warrant such matching practices.

This study tested whether matching reading instruction to reading styles positively affected reading gains compared with comparable controls in a field setting.

Frankie Oglesby is Chair of the Education Division and Director of Women's Studies at Judson College in Marion, AL. W. Newton Suter is an Associate Professor in the Department of Educational Leadership at the University of Arkansas at Little Rock. Please address correspondence regarding this paper to Dr. W. Newton Suter, Department of Educational Leadership, University of Arkansas at Little Rock, 2801 S. University Avenue, Little Rock, AR 72204.

Method

Participants

The sample included 103 third graders and 95 sixth graders (100 girls and 98 boys from 13 classes) from two city schools in the Mid-South region of the United States. The two schools were chosen because of their comparability, willingness to participate, and ease of entree. Approximately 80% of the sample was black (20% was white), and nearly 81% of the sample was identified as remedial [defined as normal curve equivalent (NCE) scores of 49 or below in the Gates-MacGinitie Reading Tests].

Instrumentation

The Reading Style Inventory (RSI) (Carbo, 1982) was used to assess students' reading styles. It is based on the learning styles described by Dunn and Dunn (1978) and provides data on the following seven dimensions: auditory, visual, tactual, kinesthetic, design, structure, and self. The inventory does not diagnose what the student does or does not know, but provides information on how the student learns best. Carbo (1988a) reported that test-retest reliabilities averaged .74. The RSI *Manual* (Carbo, 1988b) does not report traditional validity coefficients, but research cited by Carbo (1988a) may be interpreted as support for the construct validity of the RSI.

The RSI individual profiles provide teachers with the following information: (a) a description of the student's preferred learning style; (b) a recommendation of the most suitable teaching strategy, given the student's style; and (c) a listing of recommended reading materials. The RSI *Manual* (Carbo, 1988b) was also used by teachers to assist in the selection of reading methods that matched students preferred reading styles.

Reading achievement was measured by the Gates-MacGinitie Reading Tests (MacGinitie & MacGinitie, 1989) Levels 3 and 6. Form K (Total Raw Score) was used as the pretest, and Form L (Total Raw Score) was used as the posttest. The Gates-MacGinitie Reading Tests yield total raw scores ranging from 0 to 93; in the sample the pretest raw totals ranged from 10 to 81, and the posttest raw totals ranged from 17 to 89.

Design and Procedure

Random assignment of students to treatment and comparison classes was not possible; hence, 13 classes from two similar schools were selected that were matched on reading ability, aptitude, socioeconomic status, ethnic breakdown, and student-to-teacher ratio. Six classes (3 third-grade classes and 3 sixth-grade classes) were selected using matching criteria described above from a pool of volunteering teachers from one school to receive training in the use of instructional methods, strategies, and materials that were in accord with students' reading styles as measured by the RSI. Seven comparison classes (4 third-grade classes and 3 sixth-grade classes) were selected using the previously described matching criteria from a pool of volunteering teachers from the second comparable school, but were given no specific training in the matching of reading instruction and reading styles. In addition to using matching criteria, classes were also selected on the basis of teacher willingness to participate after learning about the research in an announcement sent to teachers at the schools.

Teachers in the comparison classes taught reading using a structured skills approach adopted district wide (a mastery-oriented, standard basal-series approach emphasizing decoding/phonics, word attack skills, worksheet competency and the like). (This structured, basal-series approach was the method which would have been used by the teachers of the treatment classes had they not participated in the research.) Teachers of the treatment classes were individually trained by one investigator (first author) to use a variety of teaching methods appropriate for students with varying reading styles as outlined by Carbo (1988b). These included phonic, linguistic, Orton-Gillingham, whole-word, individualized, language-experience, Fernald, and the Carbo Recorded-Book Method. Each treatment classroom prominently displayed a color-coded "Reading Style Recommendation Chart" which teachers (and students) could easily consult in order to guide reading instruction. The computerized interpretation of the RSI provides an individual profile for teachers' use in adapting instruction to style. The three-page profile identifies the child's reading style and suggests suitable strategies, methods, and materials. The RSI *Manual* (Carbo, 1988b) was provided for each teacher, and every effort was made to encourage teachers to use the profile in a manner described by Carbo (1988b). The interested reader should consult the RSI Manual which is rich in ideas for accommodating student diversity.

To assure that teachers of the treatment classes were, in fact, using accommodating techniques recommended by Carbo (1988b) and that a diversity of matching-to-style strategies was properly implemented, one investigator (first author) monitored each classroom daily. These classroom visits made it possible to answer teachers' questions and provide additional training (fine-tuning) when needed. The idea of visiting comparison classes as a control for the Hawthorne effect was not practical. Also, recent work suggests that Hawthorne controls are of limited utility and that there is little evidence in the literature to support an overall Hawthorne effect (Adair, Sharpe, & Huynh, 1989).

Both treatment and comparison classes were administered pretests to measure baseline reading ability at the beginning of the school year. The RSI scores were also collected from the treatment classes at that time. Posttest reading scores were collected from all classes 6 months later.

The quasi-experimental design for analysis was a 2 x 2 x 2 x 2 factorial with the following factors: Group (treatment and comparison), Grade (third and sixth), Sex, and Reader (remedial and developmental). The data were analyzed using the class mean and individual student as the unit of analysis. (The class mean is considered the most appropriate unit of analysis since *teachers* received the experimental training.)

Results

Initial Group Comparability

Because classes could not be assigned randomly to treatment and comparison conditions, initial reading comparability was tested with the analysis of variance on raw pretest reading scores using the factorial design previously described. The reading difference between treatment and comparison classes on the pretest was not significant, $F(1, 182) = 1.29$, $MSE = 131.61$, $p = .29$. Further, none of the two-way and higher-order interactions involving the Group factor approached statistical significance at the traditional .05 level.

Reading Styles

A detailed description of reading styles in the sample is beyond the scope of this paper, but it is clear that there was great diversity in reading style preferences. The RSI perceptual modality, for example, revealed approximately an even split between excellent/good and fair/poor auditory and visual preferences. About 43% of the sample had strong/moderate tactual preferences (opposed to mild/none), and about 66% of the sample had strong/moderate kinesthetic preferences (as opposed to mild/none). With regard to design, structure, and sociological stimuli, the sample generally preferred highly organized design, many choices in structure, and contexts which permit reading alone.

Treatment Effect: Class as Unit of Analysis

The influence of reading instruction on reading scores using the class mean as a unit of analysis was assessed in two ways. The raw gain from pretest to posttest for the six treatment classes was 10.65 ($SD = 4.40$). The corresponding raw gain for the seven comparison classes was 6.14 ($SD = 3.15$). The difference in gain was marginally significant, $t(11) = 2.15$, $p < .06$. The differential gain was also assessed with the analysis of covariance on raw posttest mean scores using the raw pretest mean scores as a covariate. This analysis also revealed higher gain for the treatment classes, $F(1, 10) = 5.47$, $MSE = 13.81$, $p < .05$. Table 1 presents the class means that were used for this analysis.

Table 1
Pretest and Posttest Class Mean Reading Scores

Group	Class	n	Testing Pretest	 Posttest
Treatment				
	1	13	44.08	49.38
	2	19	47.89	60.11
	3	21	49.43	55.57
	4	21	42.24	58.95
	5	17	39.88	49.88
	6	14	32.57	46.07
Comparison				
	1	11	37.09	40.27
	2	13	50.38	52.38
	3	14	27.50	35.86
	4	13	49.23	59.69
	5	10	44.80	49.00
	6	16	35.06	43.69
	7	16	39.00	45.13

Treatment Effect: Student as Unit of Analysis

The influence of matching reading instruction to students' reading styles using the student as the unit of analysis was tested with a 2 x 2 x 2 x 2 factorial ANCOVA using raw pretest reading scores as a covariate. Neither the Grade, Sex, nor Reader main effects approached significance. By contrast, the Group main effect reflecting the treatment-comparison difference overall was statistically significant and in the predicted direction, $F(1, 181) = 13.98$, $MSE = 87.31$, $p < .001$. Mean pretest and posttest reading scores (and standard deviations) for the Group factor are shown in Table 2. Adjusted raw posttest means were 52.84 and 47.82 for the treatment and comparison groups, respectively. The corresponding effect sizes were .47 and .30 for the unadjusted and adjusted mean differences, respectively. Using the lower effect size estimate (.30), one can conclude that the average of the treatment classes corresponds to about a percentile rank of 62 in the untreated comparison classes. None of the two-way interactions in the factorial analysis approached statistical significance.

Table 2
Pretest and Posttest Reading Scores by Group

Group	*n*	*M*	*SD*
Treatment			
Pretest	105	43.26	17.66
Posttest	105	54.11	17.09
Comparison			
Pretest	93	40.01	16.05
Posttest	93	46.38	16.52

The treatment effect was assessed also by computing reading gain scores (raw posttest minus raw pretest) and testing this difference with the analysis of variance. The treatment mean gain was 10.86 and the comparison gain was 6.37. This difference was statistically significant, $F(1, 182) = 12.12$, $MSE = 89.52$, $p < .001$, supporting the ANCOVA results presented above.

Discussion

Before discussing the implications of these findings, one must be reminded of the dangers of overinterpretation. These findings may not generalize beyond the specific characteristics of the sample (only 13 classes) nor to other measures of learning styles and reading achievement. Nonrandomized quasi-experimental research designs, such as the one used in this study, are weak with regard to ferreting out causal mechanisms. The limitations imposed by a lack of random assignment of students are compounded by the nonrandom assignment of teachers to classes. Because of these confoundings and influences such as the Hawthorne effect, it is possible that treatment classes may have achieved more than comparison classes without the matching-to-style strategies. Nevertheless, we believe the data are sufficiently strong to warrant some tentative conclusions.

This research supported the hypothesis that when instructional methods and materials are matched to identifiable reading styles, students' reading achievement scores increase more than the scores of students who are not taught with matching-to-style strategies. This suggests the need to assess students' reading style preferences and devise interventions that are compatible with specific preferences.

Teachers in this study who incorporated students' reading styles into their instructional strategies did *not* face an insurmountable task, yet their students gained more in reading achievement compared to controls. Several recommendations follow from this finding. Teachers should be encouraged to assess students' reading and learning styles in ways that convince them of their usefulness. They must be reminded that no *one* method is best for all students (Cronbach & Snow, 1977). They should experiment with instructional strategies that accommodate learning preferences in ways that maximize students' reading competencies and consequently heighten their reading enjoyment. These professionals should be encouraged to approach the art of teaching from a scientific base that incorporates the rich diversity that they undoubtedly recognize in their daily encounters with students. This recommendation is in accord with the call for teachers to evaluate their own practice (Cochran-Smith & Lytle, 1990) within the teacher-as-researcher movement (Newkirk, 1992).

It is *not* controversial that most people seem to have preferred ways of learning and that no single instructional method provides the optimal learning environment for all students. The task of the effective classroom teacher involves determining how the strengths of the student's learning and reading styles can be utilized. This requires reflective practice (Wellington, 1991). Woolfolk (1993) also reminds us that the important implication of the large body of aptitude-treatment interaction (ATI) research is *flexibility*. She stated, "When students are having difficulty learning with one teaching approach, it makes sense to try something else. If you can form some hypotheses about your students' particular needs and the reasons your current approach does not fit these needs, you should be able to find a better alternative" (p. 495).

The issue faced by educators as a whole is to what degree a system should adapt to the preferences of individual students and the degree to which students should be forced to adapt to the preferences of the system. This is more critical than ever before since young students represent an unprecedented mix of cultures and socioeconomic backgrounds. As we evaluate our present methods of teaching reading, the concept of reading styles may help educators better understand the unique way that each child learns. Carbo (1988a) summarized the teaching of reading in the following way: "The particular method by which a child learns to read is unimportant. What is important is that a child *does* learn to read with ease, enjoyment, and a high degree of competence--and as quickly as possible" (p. 326). One promising approach toward this goal involves reading style assessment and the use of matching-to-style teaching strategies.

References

Adair, J. B., Sharpe, D., & Huynh, C. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research, 59*, 215-228.

Bennett, C. (1979). The importance of cultural perspective. *Educational Leadership, 36*, 259-268.

Carbo, M. (1980). Reading style: Diagnosis, evaluation, prescription. *Academic Therapy, 16*, 45-52.

Carbo, M. (1982). *Reading style inventory (RSI)*. Roslyn Heights, NY: Learning Research Associates.

Carbo, M. (1984). Research in learning style and reading. *Theory Into Practice, 23*, 72-76.

Carbo, M. (1988a, December). The evidence supporting reading styles. *Phi Delta Kappan, 70*, 323-327.

Carbo, M. (1988b). *Reading style inventory manual*. Roslyn Heights, NY: Learning Research Associates.

Carbo, M. (1990, October). Igniting the literacy revolution through reading styles. *Educational Leadership, 47*, 26-29.

Cochran-Smith, M., & Lytle, S. L. (1990). Research on teaching and teacher research: The issues that divide. *Educational Researcher, 19*(2), 2-11.

Cronbach, L., & Snow, R. (1977). *Aptitudes and instructional methods*. New York: Irvington.

Dunn, R. (1988). Teaching students through their perceptual strengths or preferences. *Journal of Reading, 31*, 304-309.

Dunn, R., & Dunn, K. (1978). *Teaching students through their individual learning styles: A practical approach*. Reston, VA: Reston Publishing.

Koons, C. I. (1977, May). Nonpromotion: A dead end road. *Phi Delta Kappan, 59*, 701-702.

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates-MacGinitie Reading Tests*. Chicago: Riverside.

Newkirk, T. (Ed.). (1992). *Workshop by and for teachers: The teacher as researcher*. Portsmouth, NH: Heinemann.

O'Neil, J. (1990, October). Findings of styles research murky at best. *Educational Leadership, 47*, 7.

Stahl, S. A. (1988, December). Is there evidence to support matching reading styles and initial reading methods? *Phi Delta Kappan, 70*, 317-322.

Wellington, B. (1991). The promise of reflective practice. *Educational Leadership, 48*(6), 4-5.

Woolfolk, A. (1993). *Educational psychology* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 17-26

Internalizing/Externalizing Symptomatology in Subtypes of Attention-Deficit Disorder

Jose J. Gonzalez
University of Georgia

George W. Hynd
University of Georgia

Differentially diagnosing the DSM III categories of Attention-Deficit Disorder with Hyperactivity (ADD/H) and without Hyperactivity (ADD/WO) has been the focus of much debate since their introduction. This study examined the issue of whether or not children diagnosed as ADD/H or ADD/WO can be distinguished on affective measures, using an externalizing/internalizing continuum. The researchers examined 28 ADD/H and 20 ADD/WO children. When ADD/H and ADD/WO are compared on parent and teacher ratings of behavior, both informants reported greater externalizing symptomatology in ADD/H children. Without co-occurring Conduct Disorder (CD) or Oppositional Defiant Disorder (ODD) diagnosis, the parents differentiated the two groups based on internalizing symptoms, while the teachers still differentiated on externalizing criteria. These findings differentiate the two ADD subtypes into a more externalizing dimension (ADD/H with and without CD/ODD) at school/home and a more internalizing dimension (ADD/WO without CD/ODD) at home. It also confirms the notion that ADD children with a codiagnosis of CD/ODD appear to have a variety of both externalizing and internalizing problems that may confound differences between clinic samples or subtypes.

An estimated 1 to 20% of school-aged children are considered "hyperactive" (Barkley, 1990; Ross & Ross, 1982; Szatmari, Offord, & Boyle, 1989). The possible high incidence of this disorder reflects an emphasis on "hyperactivity" or Attention-Deficit Hyperactivity Disorder (ADHD) in recent years, making hyperactivity the most well-studied childhood psychiatric/psychological disorder in the last decade (Barkley, 1990). Different definitions, diagnoses, and symptoms have been described as being representative of this group. The present *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition) (*DSM-IV*) (American Psychiatric Association (APA), 1994) criteria differentiate ADHD into three categories: Attention-Deficit/Hyperactivity Disorder, Combined Type; Attention-Deficit/Hyperactivity Disorder, Predominately Inattentive Type; and Attention-Deficit/Hyperactivity Disorder, Predominately Hyperactive-Impulsive Type. The previous categorization system in *DSM-III-R* (American Psychiatric Association (APA), 1987) of one unitary ADHD syndrome led to much controversy (Goodyear & Hynd, 1992). Since 1980, and prior to *DSM-IV*, a number of studies examined the validity of the previous *DSM-III* (APA, 1980) classification typology (e.g. Lahey, Schaughency, Frame, & Strauss, 1985; Lahey, Schaughency, Hynd, Carlson, & Nieves, 1987).

This study investigates whether children diagnosed as Attention-Deficit Disorder with Hyperactivity (ADHD/H) or Attention-Deficit Disorder without Hyperactivity (ADHD/WO) can be distinguished. The *DSM-III* criteria for ADHD subtypes was used because the *DSM-IV* had not yet been published, yet the two *DSM-III* subtypes ADHD/H and ADHD/WO appear to be analogous to the *DSM-IV* ADHD: Combined Type and ADHD: Predominately Inattentive Type (McBurnett, Lahey, & Pfiffner, 1993). These subtypes were studied with particular emphasis on how internalizing or externalizing (see Achenbach, 1982 for further review of these concepts) symptoms may be differentially associated with the ADHD subtypes (e.g. Edelbrock, Costello, & Kessler, 1984; Lahey et al., 1985).

This study is based on findings by Hern (1990) and Lahey et al. (1987) that characterized ADHD/H children as having more externalizing behaviors than ADHD/WO children, who were characterized as exhibiting more internalizing behaviors. These studies also indicated that affective symptoms such as depression and anxiety were

---

Jose J. Gonzalez is a Post-Doctoral Fellow in Neuropsychology at the Neuropsychiatric Institute at the University of California, Los Angeles. George W. Hynd is a Research Professor and Clinic Director of the Center for Clinical and Developmental Neuropsychology at the University of Georgia. He is also Clinical Professor in the Department of Pediatric Neurology at the Medical College of Georgia. Reprint requests should be addressed to Jose J. Gonzalez, Ph.D., UCLA, 760 Westwood Plaza (C8-747/NPI), Los Angeles, CA 90024-1759.

more characteristic of children with ADHD/WO than for children with ADHD/H. The present study extends this line of investigation to specific characteristics of internalizing disorders, like social withdrawal which has been shown to be highly correlated with depression (see Brulle & McIntyre, 1982, for a review) and may be related to the ADHD/WO disorder (Edelbrock et al., 1984; Lahey, Schaughency, Strauss, & Frame, 1984).

Although the diagnosis of ADHD has been associated with that of depression, to date no research study has focused on differentiation of the subtypes according to depressive symptomatology. Lahey et al. (1984), using the Revised Behavior Problem Checklist (RBPC; Quay, 1983; Quay & Peterson, 1983), found significantly higher rating on the Anxiety-Withdrawal scale for ADHD/WO children than control children. The ADHD/H children could not be differentiated from controls on this measure. On the other hand, Edelbrock et al. (1984) reported that teachers rated ADHD/WO children as having more problems associated with social withdrawal and were less happy than ADHD/H children. In another study, Lahey et al. (1985) found that teachers rated ADHD/WO children as being more sluggish and slow than ADHD/H children. When considering the results of the Lahey et al. (1985) study and previous research findings (Edelbrock et al., 1984; King & Young, 1982; Lahey et al., 1984), one can categorize the ADHD group into two subtypes. One group, the ADHD/H children, was characterized by active, impulsive, and aggressive behavior; and the ADHD/WO group was noted to be more anxious, day-dreamy, lethargic, withdrawn, sluggish, more passive (yet not significant), and drowsy (Edelbrock et al., 1984; Lahey et al., 1985; Lahey et al., 1987; Lahey et al., 1984).

A comparison of behavioral characteristics of children diagnosed clinically as ADHD/H and ADHD/WO was reported by Lahey et al. (1987). They found that [illegible]% of the ADHD/WO children had a codiagnosis of an internalizing disorder (anxiety or depression), while only 10% of the ADHD/H children received such codiagnosis. However, these results were reported when the analyses were repeated after eliminating children from both groups who had a codiagnosis of Conduct Disorder. In the original analyses, the groups did not differ on ratings of anxiety or depression.

In a more recent study, Barkley, DuPaul, and McMurray (1990) reported that 41% of the ADHD/H group met criteria for Oppositional Defiant Disorder and more than 21% met criteria for Conduct Disorder, while the ADHD/WO group were 19 and 6%, respectively. Although there was a low rate of Major Depressive Disorder, the ADHD/WO group had significantly more of these symptoms endorsed than the ADHD/H. It appears that the ADHD/H children, in general, were more likely to manifest symptoms of other disruptive disorders than the ADHD/WO. Also, Barkley et al. (1990) suggests that both at home and at school ADHD/H children have more pervasive conduct problems, are more impulsive and hyperactive, and are more aggressive and delinquent than ADHD/WO children. Both of these groups appeared to have a greater diversity of internalizing-externalizing symptoms than the Learning Disability (LD) and normal control groups.

It is unclear, according to these studies, what the relationship is between the ADHD subtypes and actual affective symptoms. Some of these studies have relied predominately on teacher ratings as criteria for defining their groups and then in assessing differences between groups. Some of these studies (e.g. Lahey et al., 1984; Lahey et al., 1987) may have confounded their independent and dependent variables, in which case differences on the teacher ratings, serving as the dependent measures, were subsequently found. Many of these studies have also used nonclinical samples; therefore, the extent to which their results are representative of clinic-referred children with potentially more serious ADHD symptoms is questionable. Nevertheless, according to the research to date, one could expect marked differences in behavior patterns for the two ADHD subtypes. One subtype is characterized by hyperactivity and more behavior/conduct problems associated with externalizing disorders, while the other subtype is characterized by less motor overactivity and more anxious, shy, and withdrawn behaviors associated with internalizing disorders (Dykman & Ackerman, 1993; Lahey et al., 1987). It was then hypothesized that children with ADHD/H would manifest more externalizing disorders/symptoms and children with ADHD/WO would manifest more internalizing disorders/symptoms.

Method

Subjects

Subjects were 48 children referred to the Center for Clinical and Developmental Neuropsychology (CCDN), an outpatient diagnostic and referral clinic. Subjects had a mean age of 10 years, 4 months (*SD* = 32.22 months). Socioeconomic status (SES), based on parental education, for this group ranged across all levels of SES, but most were from low middle to middle SES. Current grade placements ranged from kindergarten to 12th grade. The subjects were predominately Caucasian (45 Caucasian-Americans, 2 African-Americans, and 1 Hispanic-American), and male (34 males, 14 females). The identified sample was composed of 28 children with ADHD/H and 20 children with ADHD/WO from low middle to middle SES families where no parent psychopathology was present. The CCDN serves children and adolescents of all ages, although this sample ranges from ages 6 to 16. The children were taken from consecutive referrals to the clinic over a 3-year period (1987-1990). First, subjects that showed evidence of overt neurological disorder (e.g., epilepsy, tic disorder, Tourette's), psychotic disorder, or mental retardation (Full Scale Standard Score on the WISC-R < 70) were eliminated from analysis. Then, all children whose primary diagnoses was Attention-Deficit/Hyperactivity Disorder were over the age of 6. Also, two children with a secondary diagnosis of Attention-Deficit/Hyperactivity Disorder and a primary diagnosis of Learning Disabilities were added because of their codiagnosis of ADHD/WO.

The purpose of the CCDN is to provide a clinical service and to conduct pertinent research; therefore, written informed consent of the parent and the oral consent of the child were required in every case. Referral sources were primarily from area physicians, other clinical services, and schools.

Instruments

Instrumentation chosen for analyses in this study were selected from the comprehensive diagnostic battery administered as part of the neuropsychological evaluation conducted on each subject. Previous research (Achenbach, 1982; Achenbach & Edelbrock, 1979; Edelbrock et al., 1984; Hynd et al., 1991; Lahey et al., 1985; Lahey et al., 1987) that used the measures chosen for this study indicated their utility in differentiating children with ADHD/H, ADHD/WO, and other clinic-referred children on internalizing and externalizing symptoms.

Instrumentation (dependent measures) included the Achenbach Child Behavior Checklist--Teacher's Report Form (CBCL-T) and the Parent Form (CBCL-P), from which the items on the Internalizing and Externalizing scales were used as variables. The CBCL-P is designed to report the behavioral problems and competencies of children ages 4 through 16, as reported by their parents. For the different age groups, 4-5, 6-11, and 12-16, the items were factor analytically derived to comprise the different scales. The Internalizing and Externalizing scales were developed by using the items that loaded the highest on either scale. The test-retest reliability for the Internalizing/Externalizing scale scores across ages ranged from .81 to .97 (Achenbach & Edelbrock, 1983).

The CBCL-T was designed to obtain teacher reports on the student's behavioral problems and adaptive functioning skills. The test-retest reliability for special education students on the Internalizing/Externalizing scales ranges from .82 to .92 across ages.

The Structured Interview for Diagnostic Assessment of Children (SIDAC) (all sections excluding the ADHD section were used), reported by the biological mother, was also used. Using the SIDAC, the symptoms endorsed by the mother in each disorder were differentiated as either being externalizing or internalizing disorders/symptoms. The use of these instruments was chosen to assess ADHD/H and ADHD/WO children on an internalizing-externalizing continuum, avoiding the problem found in previous research (Achenbach, 1982; Edelbrock et al., 1984; Hern, 1990; Lahey et al., 1984; Lahey et al., 1987), where categorical diagnosis of externalizing or internalizing disorders was made without examining the actual symptomatology associated with the two dimensions of behavior.

The dependent measures/variables were not used to define the experimental groups nor were they used to diagnose any of the experimental groups.

Procedure

Each subject and at least one biological parent, usually the mother, participated in a comprehensive, day-long neuropsychological evaluation. Parent interviews consisted of the Structured Interview for Diagnostic Assessment of Children (SIDAC), a modified and updated version of the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS; Puig-Antich & Chambers, 1978) to include all symptoms used in the *DSM-III* (APA, 1980) and *DSM-III-R* (APA, 1987) for the following diagnostic categories: Major Depressive Episode (MDE), Dysthymic Disorder (DD), Attention-Deficit Disorder With and Without Hyperactivity (ADHD/H and ADHD/WO), Conduct Disorder (CD), Separation Anxiety Disorder (SAD), Overanxious Disorder (OAD), Oppositional Defiant Disorder (ODD), and Psychosis. The parents also completed the Child Behavior Checklist Parent's Report Form (CBCL-P: Achenbach & Edelbrock, 1983). In addition to the parent's rating on the CBCL, each child's principal classroom teacher was contacted and completed the CBCL-Teacher's Report Form (CBCL-T: Achenbach & Edelbrock, 1986). Both the parent and the teacher completed the SNAP (Pelham & Murphy, 1981) and the Personality Inventory for Children-Revised (PIC-R) (Wirt, Seat, Broen, & Lachar, 1981).

Diagnosis was based on procedures employed in other clinics where similar measures have been administered (Hynd et al., 1991; Lahey et al., 1987). Combining all these sources of information and using a global diagnostic procedure, including interviews, observations (school/test taking), and standardized tests (excluding the CBCL-P and CBCL-T), decisions were made regarding the presence or absence of the *DSM-III* and *DSM-III-R* symptoms. This resulted in a comprehensive diagnostic classification of each child. To insure reliability of the clinical diagnosis, a second clinician involved in the case gave an independent diagnosis of the child based on the available information and instruments administered, as well as observation of the child, excluding the rating scales (CBCL-P and CBCL-T). The final *DSM-III* and/or *DSM-III-R* diagnoses reflected interclinician resolution and were used for diagnosing the children into ADHD/H and ADHD/WO for all analyses. Interclinician reliability using this procedure was 77.3% (Kappa = .71), with 79% agreement for the ADHD/H and 80% agreement for the ADHD/WO diagnosis, equivalent to reliabilities from similar procedures at other clinics (Lahey et al., 1987). The first clinician was an advanced graduate student in school psychology, while the second was the licensed faculty member who directed the Clinic. Neither clinician involved in the diagnosis was informed of the purpose of the study.

Clinicians formed two experimental groups on the basis of *DSM-III* diagnosis, based on the SIDAC (the ADHD section only) and the SNAP data. Group 1 was composed of 28 children with a clinical diagnosis of ADHD/H, while Group 2 consisted of 20 children with a diagnosis of ADHD/WO. No control group was used because this investigation was only interested in the comparisons between ADHD/H and ADHD/WO. It should be noted that both groups contained children with a sole diagnosis of ADHD/H and ADHD/WO as well as children with co-diagnoses of other disorders, but these were always secondary to the ADHD diagnosis. Table 1 provides descriptive and frequency data for demographic variables and background variables.

Analyses

After generating descriptive statistics and conducting one-way ANOVAs to ascertain group equivalence on potential mediating or moderator variables such as age or child's intellectual functioning, a *t*-test was calculated on the *t* scores of the Teacher and Parent Child Behavior Checklists for the Internalizing and Externalizing scales to evaluate differences between each group. Differentiation through symptom analysis of affective behavior was conducted. On the SIDAC, a Chi Square was performed for each of the items in the four internalizing and two externalizing disorders for each group in order to assess differences in internalizing and externalizing symptomatology. Due to the possible chance factors associated with making many simultaneous comparisons, all probability values less than .01 were considered "clearly" significant and those less than .05 were considered "tentatively" significant.

Table 1
Demographic and background characteristics
of all subject groups.

	ADHD/H	ADHD/WO
Total Sample Size (N)	28	20
Gender		
Males	21	13
Females	7	7
Codiagnoses		
Average Number	1.00	1.20
Frequency of *DSM-III-R* Diagnoses		
Developmental Arithmetic Disorder	4	9
Conduct Disorder	7	0
Oppositional Defiant Disorder	6	1
Developmental Reading Disorder	3	3
Dysthymic Disorder	1	2
Major Depression	2	1
Separation Anxiety Disorder	2	1
Developmental Expressive Writing Disorder	1	1
Avoidance Disorder	0	1
Developmental Articulation Disorder	0	1
School Phobia	1	0

	Mean	SD	Mean	SD
Chronological Age Months	121.6	32.5	134.5	29.9
IQ				
VIQ	102.7	16.0	104.9	17.2
PIQ	102.5	18.6	97.3	13.1
FSIQ	102.8	17.1	101.3	14.6

Results

The primary purpose of the present study was to determine whether there were differences between children diagnosed as ADHD/H and children diagnosed as ADHD/WO, according to affective symptomatology, and, subsequently, the affective symptoms associated with each group. The analysis included the items on the SIDAC and the CBCL-P Internalizing/Externalizing scales for parent ratings; and for teacher ratings, the CBCL-T Internalizing/Externalizing scales.

Data Analysis

Preliminary analyses were conducted in order to examine age and FSIQ differences among groups. Both variable comparisons, using *t*-tests with a pooled estimate of variance, yielded insignificant mean differences. As neither age nor IQ differed, these variables were dropped from consideration in further analyses.

Planned comparisons with one-tailed *t*-tests were used for all measures that had scaled scores and Chi Square tests were used for the comparison of the frequency of endorsed individual items. It was expected that the ADHD/H group would have greater mean scores on the externalizing/items scales and the ADHD/WO would have greater mean scores on the internalizing scales/items. All measures were not available for every child; thus, group sizes varied for each comparison. Separate estimates of variance were utilized in each case to control for unequal group sizes.

Comparisons on Parent Reported Measures. On the behavior rating scales, significant results were found in the planned comparisons in many of the maternal reported measures. The ADHD/H group had higher mean scores than the ADHD/WO group on the Externalizing factor scale of the CBCL-P and the CBCL-T (see Table 2). As predicted, the ADHD/H group had significantly greater mean scores on the parent-reported Externalizing factor (68.54) than did the ADHD/WO group (61.21, $p<.01$). The parent-reported Internalizing factor comparison was not significant.

Table 2
Group means, degrees of freedom, *p* values, and standard deviations for parent and teacher reported measures.

	ADHD/H		ADHD/WO				
	N	*X* (SD)	*N*	*X* (SD)	*t-test*	*df*	*p value*
CBCL-P (INT.)	28	62.39 (11.51)	19	63.68 (9.02)	-0.41	45	0.68
CBCL-P (EXT.)	28	68.54 (10.09)	19	61.21 (6.67)	2.78	45	0.01
CBCL-T (INT.)	25	58.88 (8.86)	15	63.27 (9.05)	-1.50	38	0.14
CBCL-T (EXT.)	25	65.88 (7.11)	15	60.73 (4.45)	2.52	38	0.02

The group means are *t*-scores with a mean of 50 and a standard deviation of 10. *p*-values of .0125 or less are statistically significant after applying the Bonferroni correction.

An analysis of the SIDAC items/symptoms showed that the ADHD/H group had significantly higher mean frequency scores than the ADHD/WO group on externalizing items/symptoms, while the opposite was true for the internalizing items/symptoms. See Table 3 for group comparisons on the SIDAC items. Only the significant items are noted. As expected, *the ADHD/H group* "frequently gets in trouble at home or at school for breaking important rules" more than the ADHD/WO group ($p<.01$). They also "tell lies quite a bit more often than boys/girls their own age ($p<.05$), steal outside the home ($p<.05$), bully other children regularly ($p<.05$), often get into physical fights ($p<.05$), have been physically cruel to animals ($p<.05$), and have deliberately damaged or destroyed other people's property ($p<.05$)" tentatively more than the ADHD/WO children. On the other hand, the *ADHD/WO group* was reported to have "complained of headaches, stomachaches, nausea, or vomiting on many school days--much more than on weekends or holidays--or at other times when he/she had to be away from his/her mother" tentatively more than the ADHD/H group, ($p<.05$). Parents also reported that ADHD/H children "often deliberately do things that annoy other people, like grabbing other children's belongings" ($p<.01$). They also "often actively defy or refuse to comply with requests or rules, like refusing to do chores at home" ($p<.05$), and are "often spiteful or vindictive" ($p<.05$) tentatively more than the ADHD/WO group.

Analyses without the CD/ODD codiagnosis. In the Lahey et al. (1987) study, they report that ADHD/WO children were rated as having a greater degree of anxiety and depressive symptomatology than ADHD/H children only when those children with a codiagnosis of Conduct Disorder were removed from the data pool. Also, behavioral and laboratory reports show that the mixed group of ADHD/H plus CD and the "pure" ADHD/H group behave as separate groups across different variables/dimensions (Forness, Swanson, Cantwell, Youpa, & Hanna, 1992; Lahey et al., 1987; Shaywitz & Shaywitz, 1988; Szatmari et al., 1989). Further analyses were therefore conducted in which children with a codiagnosis of Conduct Disorder or Oppositional Defiant Disorder were removed from the two groups. Group sizes were then 15 for the ADHD/H group and 18 for the ADHD/WO group. As stated before, all measures were not available for each child, and group size varied per comparison. See Table 4 for group means, degrees of freedom, and *p* values.

Table 3
Percentage of responses on SIDAC items, sample sizes, and p values.

Cat	Item	Symptom	ADHD/H %	ADHD/H N	ADHD/WO %	ADHD/WO N	X^2	*p value*
CD	1	Frequently gets in trouble for breaking rules	37.50	28	6.25	20	11.52	0.01
CD	4	Tells lies	20.83	28	4.17	20	4.11	0.04
CD	6	Steals	10.42	28	0.00	20	3.99	0.05
CD	9	Bullies others	14.58	28	0.00	20	5.85	0.02
CD	10	Fights	16.67	28	2.08	20	4.26	0.04
CD	13	Physically cruel to animals	10.42	28	0.00	20	3.99	0.05
CD	14	Damages other's property	12.50	28	0.00	20	4.90	0.03
SAD	7	Somatic complaints when away from parent	8.33	28	16.67	20	4.11	0.04
ODD	3	Actively defies rules	29.73	23	5.41	14	4.30	0.04
ODD	4	Does things that deliberately annoy others	29.73	23	0.00	14	9.53	<0.005
ODD	8	Often spiteful or vindicative	18.92	23	0.00	14	5.26	0.02

Cat = Category. CD = Conduct Disorder. SAD = Separation Anxiety Disorder. ODD = Oppositional Defiant Disorder.
% = Percentage of items in the cell. *p*-values of .0045 or less are statistically significant after applying the Bonferroni correction.

The results of the parent-reported behavior ratings, as predicted, demonstrate that the ADHD/WO group has a tentatively higher mean score on the Internalizing factor (64.11) in comparison to the ADHD/H group, minus CD and ODD children, (55.60, $p<.05$). On the Externalizing factor, the ADHD/H group essentially achieved the same mean score (61.53) as did the ADHD/WO group (61.39).

Table 4
Group means, degrees of freedom, p-values, and standard deviations for parent and teacher reported measures of ADHD children without a codiagnosis of CD or ODD.

	ADHD/H *N*	ADHD/H *X* (SD)	ADHD/WO *N*	ADHD/WO *X* (SD)	*t-test*	*df*	*p value*
CBCL-P (INT.)	15	55.60 (10.03)	18	64.11 (9.08)	-2.56	31	0.02
CBCL-P (EXT.)	15	61.53 (7.06)	18	61.39 (6.82)	0.06	31	0.95
CBCL-T (INT.)	14	59.14 (8.37)	14	63.57 (9.31)	-1.32	26	0.20
CBCL-T (EXT.)	14	64.50 (2.85)	14	61.36 (3.88)	2.45	26	0.02

The group means are *t*-scores with a mean of 50 and a standard deviation of 10. *p*-values of .0125 or less are statistically significant after applying the Bonferroni correction.

BEST COPY AVAILABLE

On the SIDAC, only those items that corresponded to internalizing items/symptoms differentiated the two groups. See Table 5 for group comparisons of the SIDAC items. Again, only the significant items are noted. Mothers reported that the ADHD/WO children more frequently than ADHD/H children "complained of headaches, stomachaches, nausea, or vomiting on many school days--much more than on weekends or holidays--or at other times when he/she had to be away from his/her mother" ($p<.01$). They also "really didn't want to go to school or refused to go to school, and instead wanted to stay with his/her mother" ($p<.05$) more than the ADHD/H children.

Table 5
Percentage of responses on SIDAC items, sample sizes, and *p*- values of children with ADHD and without a codiagnosis of CD or ODD.

Cat	Item	Symptom	ADHD/H %	ADHD/H *N*	ADHD/WO %	ADHD/WO *N*	X^2	*p value*
SAD	3	Refuses to go to school and wants to stay with parent	0.00	15	14.71	19	4.63	0.03
SAD	7	Somatic complaints when away from parent	0.00	15	23.53	19	8.26	<.005

Cat = Category. SAD = Separation Anxiety Disorder.
% = Percentage of items in the cell. *p*-values of .025 or less are statistically significant after applying the Bonferroni correction.

Comparisons on teacher-reported measures. On the behavior rating scales, the ADHD/H group had higher mean scores than the ADHD/WO group on the Externalizing factor scale of the CBCL-T (see Table 2). As predicted, the ADHD/H group was reported to be (tentatively) significantly higher on the teacher-reported Externalizing factor ($M = 65.88$) than was the ADHD/WO group ($M = 60.73$, $p<.05$). Although the teacher-reported Internalizing factor comparison did not reach significance, the ADHD/WO group was reported to have a higher mean score (63.27) than the ADHD/H group (58.88).

Analyses without the CD/ODD codiagnosis. When the subjects with a codiagnosis of Conduct Disorder and/or Oppositional Defiant Disorder were removed from the data pool, there were 14 ADHD/H and 14 ADHD/WO children left, with a mean age and range similar to the first sample. See Table 5 for group means, degrees of freedom, and p values. The ADHD/H children were reported to have a tentatively higher mean score on the Externalizing factor (64.50) than the ADHD/WO group (61.36, $p<.025$). In addition, although the ADHD/WO group had a higher mean score on the Internalizing factor (63.57) than the ADHD/H group (59.14), as expected, the difference was not significant.

Discussion

Research suggests that affective symptomatology is associated with ADHD in children. Overall, however, the results from this study suggest that differences exist between children with ADHD/H and ADHD/WO in terms of affective symptomatology and that a codiagnosis of CD/ODD exerts a significant influence on the manifestation and/or endorsement of these symptoms. Children with ADHD/H were viewed as exhibiting a greater degree of externalizing symptomatology, while the ADHD/WO children were viewed as exhibiting more internalizing symptomatology.

Specifically, the findings of this study can be summarized as follows:

1. The differences found on the parent- and teacher-reported behavior ratings demonstrated that the ADHD/H group had more externalizing symptoms, as a whole, than the ADHD/WO group. This difference on the CBCL-P Externalizing scale was significant when the factor scales were compared for each group. On the CBCL-T, the Externalizing scale significantly differentiated the two groups when factor scales were compared.

2. On the SIDAC, the ADHD/H group was reported to more frequently manifest seven Conduct Disorder and three Oppositional Defiant Disorder symptoms that were externalizing in nature and differentiated them from the ADHD/WOs. On the other hand, there was one Separation Anxiety Disorder item that differentiated the ADHD/WO children from the ADHD/H children.

3. When the children with a codiagnosis of Conduct Disorder and/or Oppositional Defiant Disorder were removed from the two groups, the parents reported a significant difference on the Internalizing factor scale of the CBCL-P, with the ADHD/WOs having higher mean scores. On the CBCL-T, the ADHD/H group manifested significant differences on the Externalizing factor scale.

4. Two items on the SIDAC that were internalizing (SAD) symptoms distinguished the ADHD/WO group from the ADHD/H group.

Some investigators conclude that affective symptomatology in children with ADHD may be due to difficulties caused or induced by the disorder itself, such as discouragement caused by academic difficulties (Cantwell & Carlson, 1980). Although previous research has shown that children with ADHD/WO have more frequent academic difficulties (Carlson, Lahey, & Neeper, 1986; Edelbrock et al., 1984; Hynd et al., 1991) and receive more placements in LD classes (Barkley et al., 1990), ADHD/H children receive similar placements in reference to controls.

Another explanation for these results is that there appears to be a difference between the ADHD/H and ADHD/WO groups when the codiagnosis of CD or ODD are present (Lahey et al., 1987; Shaywitz & Shaywitz, 1988). The ADHD children with a codiagnosis of CD/ODD are reported to have a greater number of internalizing and externalizing symptomatology than the ADHD/H children who do not have CD/ODD codiagnoses. This could be due to an over-reporting of symptoms by the rater. In this case, the rater observes many behavior problems in supposedly one area, externalizing symptoms, and reports non-discriminately behavior problems in all areas. They appear to observe one disruptive behavior and over-report other behaviors. This has been examined in mothers who rate themselves as more depressed. These depressed mothers also rate their children as more deviant (Forehand, Wells, McMahon, Griest, & Rogers, 1982; Griest, Wells, & Forehand, 1979; Rickard, Forehand, Wells, Griest, & McMahon, 1981). In two of these studies (Forehand et al., 1982; Griest et al., 1979), measures of maternal depression were more strongly related to child adjustment then were external evaluations of child behavior. However, in this sample parental scores on MMPIs did not show clinically significant scale score elevations.

A more plausible explanation is that children with a codiagnosis truly have extensive behavior problems in several areas of functioning. Several studies (Loeber, 1988, 1990; Szatmari et al., 1989) reported more severe symptoms/behaviors, based on frequency, in mixed groups of ADHD/H plus CD than in pure ADHD/H or pure CD. These explanations are believed to underlie the present findings in that the ADHD/H group, which was reported to manifest externalizing symptoms, was reported to exhibit these behaviors to a lesser extent when the children with codiagnoses of CD and/or ODD were removed from the sample. Although the groups seem to overlap, there appear to be three distinct groups (Loney, 1987; Szatmari et al., 1989; Taylor, 1986; Trites & Laprade, 1983). Also, as a whole, the number of symptoms that differentiated the two groups was less when the CD and/or ODD codiagnosed children were excluded. Once the CD and/or ODD children were excluded, the ADHD/WO group displayed internalizing symptoms, as a cluster, on the parent ratings. This can possibly be explained by the parents' perceptions, because they are more susceptible to observe these types of "internal" behavior at home, as opposed to teachers. The ADHD/H group, on the other hand, exhibited externalizing symptoms as a cluster on the teacher ratings. Perhaps this is due to the teachers' reference group (other students in the class), as opposed to the parents who may be more tolerant of their own children and under-report behaviors of this type. Another explanation is that the classroom setting may be a more likely one for these externalizing behaviors to be isolated and observed, and that is why these behaviors are only seen in that context.

It appears that ADHD/H children are considerably more likely to manifest symptoms of other disruptive behaviors/disorders than are ADHD/WO children. This suggests that attentional problems may predispose children toward greater risk for these externalizing or internalizing problems, whereas the additional presence of overactivity in the presence of CD/ODD considerably increases the risk of additional externalizing symptoms and their severity in the ADHD/H group or vice versa. These results tentatively indicate that ADHD/H and ADHD/WO may be dissimilar psychiatric disorder subtypes with qualitatively different affective symptomatology.

The co-existing externalizing or internalizing problems have direct implications for designing educational interventions, determining outcomes, and eligibility. Previous studies have shown greater incidence of aggressive/disruptive behavior disorders (Barkley, et al., 1990; Cantwell & Baker, 1992) in ADHD/H with and without CD/ODD. In addition, the ADHD/H with CD/ODD appears to have a strong association with adult psychopathy and criminality (Magnusson, 1988). Children with ADHD/H, and more pervasively if they have a codiagnosis of CD/ODD, exhibit more externalizing or behavioral problems, while the ADHD/WO show more subtle emotional problems manifested at home. Most importantly, this is relevant for diagnostic and assessment issues, which suggest that a more comprehensive, multi-informer behavioral evaluation of the child will lead to better identification of ADHD subtypes. As a result, the subsequent classroom interventions can be geared more appropriately to the co-existing behavioral or emotional problems and their severity. If this subtyping can be found early in childhood, initial interventions can reduce and prevent further development of some of these externalizing symptomatologies.

Finally, although the present study has some clear-cut findings, it has some limitations. These findings are robust for the subtypes with CD/ODD included, but further analyses, excluding those codiagnoses, relied on a smaller sample. In addition, multiple comparisons were made without using a correction procedure, although the interpretation of the data was more conservative. Finally, although few studies have looked at the developmental changes in emotional/behavioral symptomatology, the groups in this study represented a wide age range (6-16 years old). These findings, taken tentatively, suggest that practitioners and educators should consider diagnostic subgrouping of ADHD. Careful evaluation of the behavioral/emotional symptoms should also be helpful not only for identification of these subgroups but also for the future design of interventions for them.

References

Achenbach, T. M. (1982). Assessment and taxonomy of children's behavior disorders. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 5, pp. 1-38). New York: Plenum Press.

Achenbach, T. M., & Edelbrock, C. S. (1979). The Child Behavior Profile: II. Boys aged 12-16 and girls aged 6-11 and 12-16. *Journal of Consulting and Clinical Psychology, 47*, 223-233.

Achenbach, T. M., & Edelbrock, C. S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Edelbrock, C. S. (1986). *Manual for the Teacher Report Form and the Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barkley, R. A. (1990). *Attention Deficit Hyperactivity Disorder: A handbook for diagnosis and treatment*. New York: Guilford Press.

Barkley, R. A., DuPaul, G. J., & McMurray, M. B. (1990). Comprehensive evaluation of Attention Deficit Disorder with and without Hyperactivity as defined by research criteria. *Journal of Consulting and Clinical Psychology, 58*, 775-789.

Brulle, A. R., & McIntyre, T. C. (1982). *Socially withdrawn children: A review*. U.S. Department of Health, Education, and Welfare.

Cantwell, D. P., & Baker, L. (1992). Attention Deficit Disorder with and without Hyperactivity: A review and comparison of matched groups. *Journal of the Academy of Child and Adolescent Psychiatry, 32*, 432-438.

Cantwell, D. P., & Carlson, C. L. (1980). *Affective disorders in childhood and adolescence: An update*. New York: Spectrum.

Carlson, C. L., Lahey, B. B., & Neeper, R. (1986). Direct assessment of the cognitive correlates of Attention Deficit Disorders with and without Hyperactivity. *Journal of Psychopathology and Behavioral Assessment, 8*, 69-86.

Dykman, R. A., & Ackerman, P. T. (1993). Behavioral subtypes of Attention Deficit Disorder. *Exceptional Children, 60*, 132-141.

Edelbrock, C. S., Costello, A. J., & Kessler, M. D. (1984). Empirical corroboration of Attention Deficit Disorder. *American Academy of Child Psychiatry, 23*, 285-290.

Forehand, R., Wells, K. C., McMahon, R. J., Griest, D., & Rogers, T. (1982). Maternal perception of maladjustment in clinic-referred children: An extension of earlier research. *Journal of Behavioral Assessment, 4*, 145-151.

Forness, S. R., Swanson, J. M., Cantwell, D. P., Youpa, D., & Hanna, G. L. (1992). Stimulant medication and reading performance: Follow-up on sustained dose in ADHD boys with and without Conduct Disorders. *Journal of Learning Disabilities, 25*, 115-123.

Goodyear, P. R., & Hynd, G. W. (1992). Attention Deficit Disorder with (ADD/H) and without (ADD/WO) Hyperactivity: Behavioral and neuropsychological differentiation. *Journal of Clinical Child Psychology, 21*, 273-305.

Griest, D., Wells, K. C., & Forehand, R. (1979). An examination of predictors of maternal perception of maladjustment in clinic-referred children. *Journal of Abnormal Psychology, 88*, 277-281.

Hern, K. L. (1990). *The relationship between affective symptomatology and the Attention Deficit Disorders*. Unpublished doctoral dissertation, University of Georgia.

Hynd, G. W., Lorys, A. R., Semrud-Clikeman, M., Nieves, N., Huettner, M., & Lahey, B. B. (1991). Attention Deficit Disorder without Hyperactivity (ADD/WO): A distinct behavioral and neurocognitive syndrome. *Journal of Child Neurology, 6* (Supplement), s37-s43.

King, C., & Young, R. D. (1982). Attentional deficits with and without hyperactivity: Teacher and peer percep-tions. *Journal of Abnormal Child Psychology, 10*, 483-496.

Lahey, B. B., Schaughency, E. A., Frame, C. L., & Strauss, C. C. (1985). Teacher ratings of attention problems in children experimentally classified as exhibiting Attention Deficit Disorders with and without Hyperactivity. *Journal of the American Academy of Child and Adolescent Psychiatry, 24*, 613-616.

Lahey, B. B., Schaughency, E. A., Hynd, G. W., Carlson, C. L., & Nieves, N. (1987). Attention Deficit Disorders with and without Hyperactivity: A comparison of behavioral characteristics of clinic referred children. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 718-723.

Lahey, B. B., Schaughency, E. A., Strauss, C. C., & Frame, C. L. (1984). Are Attention Deficit Disorders with and without Hyperactivity similar or dissimilar disorders? *Journal of the American Academy of Child and Adolescent Psychiatry, 23*, 302-309.

Loeber, R. (1988). Natural histories of conduct problems, delinquency, and associated substance use. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 11, pp. 73-124). New York: Plenum Press.

Loeber, R. (1990). Developmental and risk factors of juvenile antisocial behavior and delinquency. *Clinical Psychology Review, 10*, 1-42.

Loney, J. (1987). Hyperactivity and aggression in the diagnosis of attention deficit disorder. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 10, pp. 99-135). New York: Plenum Press.

Magnusson, D. (1988). *Individual development from an interactional perspective: A longitudinal study*. Hillsdale, NJ: Erlbaum.

McBurnett, K., Lahey, B. B., & Pfiffner, L. J. (1993). Diagnosis of Attention Deficit Disorders in DSM-IV: Scientific basis and implications for education. *Exceptional Children, 60*, 108-117.

Pelham, W. E., & Murphy, H. A. (1981). *The SNAP Checklist: A teacher checklist for identifying children with attention deficit disorders.* Unpublished manuscript.

Puig-Antich, J., & Chambers, W. (1978). *The Schedule for Affective Disorders and Schizophrenia for School-aged Children.* New York: New York State Psychiatric Institute.

Quay, H. C. (1983). A dimensional approach to behavior disorder: The Revised Behavior Problem Checklist. *School Psychology Review, 12*, 244-249.

Quay, H. C., & Peterson, D. R. (1983). *Interim manual for the Revised Behavior Problem Checklist.* Coral Gables, FL: University of Miami.

Rickard, K. M., Forehand, R., Wells, K. C., Griest, D. L., & McMahon, R. J. (1981). A comparison of mothers of clinic-referred deviant, clinic-referred nondeviant, and nonclinic children. *Behaviour Research and Therapy, 19*, 201-205.

Ross, D. M., & Ross, S. A. (1982). *Hyperactivity: Current issues, research, and theory* (2nd ed.) New York: Wiley & Sons.

Shaywitz, S. E., & Shaywitz, B. A. (1988). Attention Deficit Disorder: Current perspectives. In J. F. Kavanagh and T. J. Truss, Jr. (Eds.), *Learning Disabilities: Proceedings of the National Conference* (pp. 369-523). Parkton, MD: York Press.

Szatmari, P., Boyle, M. H., & Offord, D. R. (1989). ADDH and conduct disorder: Degree of diagnostic overlap and differences among correlates. *American Academy of Child and Adolescent Psychiatry, 28*, 865-872.

Szatmari, P., Offord, D. R., & Boyle, M. H. (1989). Ontario Child Health Study: Prevalence of Attention Deficit Disorder with Hyperactivity. *Journal of Child Psychology and Psychiatry, 30*, 219-230.

Taylor, E. A. (1986). Attention deficit. In E. A. Taylor (Ed.), *The overactive child* (pp. 73-106). Philadelphia: Lippincott.

Trites, R. L., & Laprade, K. (1983). Evidence for an independent syndrome of hyperactivity. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 24*, 573-580.

Wirt, R. D., Seat, P. D., Broen, W. E., & Lachar, D. (1981). *Personality Inventory for Children - Revised format administration booklet.* Los Angeles: Western Psychological Services.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 27-33

Reliability and Validity of Dimensions of Teacher Concern

Douglas W. Schipull, Carolyn K. Reeves, and Richard Kazelskis
University of Southern Mississippi

Ten dimensions of teacher concern were identified through a factor analysis of 568 inservice teachers' responses to the Teacher Concerns Checklist (TCCL). Test-retest and alpha coefficients were obtained for scores on each of the 10 factors. The validity of the 10 factor scores was examined by studying the relationships of the TCCL factor scores to subarea scores on the Quality of Teacher Work Life Survey (QTWLS) and the Teacher Stress Inventory (TSI). It was concluded that the TCCL factors represent important dimensions of teacher concern which can be measured with sufficient reliability and validity, providing a useful research tool for the study of teacher concerns.

One way to improve and reform educational practice is to reduce the level of stress experienced by many classroom teachers (Raschke, Dedrick, Strathe & Hawkes, 1985). The day-to-day concerns, problems, and frustrations faced by teachers often evolve into a stressful state which, if not alleviated, can lead to burnout. According to Kyriacou (1987), teacher stress may be defined as the experience of unpleasant emotions resulting from aspects of work as a teacher, and he defined teacher burnout as the syndrome resulting from prolonged teacher stress. George (1978) has stated that if teacher concerns are not identified and resolved, they will most likely lead to teacher stress and, ultimately, to teacher burnout.

It is probable that levels of teacher concern and stress vary during the school year depending on job-related events as well as individual teacher personalities. It is likely, however, that some areas of teacher concern are more pervasive and ongoing. Therefore, it is important to develop a means of identifying the more prevalent teacher concerns and frustrations associated with the work environment, so that intervention procedures can be developed and implemented before concerns lead to stress and/or burnout.

Since the identification of teacher concerns is prerequisite to the development of strategies and procedures for alleviating concerns, accurate determination of teacher concerns is necessary. In the early 1970s, Frances Fuller and Archie George developed the Teacher Concerns Checklist, Form B (TCCL) to identify teacher concerns. Initially, the TCCL yielded five scale scores; however, a factor analysis of the instrument by George (1978) resulted in an 11 factor solution. Since George was primarily interested in the 3-factor (i.e., self, task, and impact) Fuller theory of teacher concerns, 8 of the 11 factors were not utilized. Additionally, George's sample was comprised of preservice teachers, inservice teachers, and school principals. Because, theoretically, differences in concerns are anticipated between inservice and preservice teachers, George's factor analytic results may be spurious due to level differences in the groups; in fact, it is quite likely that the concerns of school principals are different from those of preservice and inservice teachers. In fact, Kazelskis and Reeves (1987) found that a factor analysis of TCCL responses of preservice teachers resulted in 12 factors which only moderately reflected the 11 dimensions found by George. At present, there has been no definitive investigation of the dimensions of concern present among inservice teachers. Results of the studies by George (1978) and Kazelskis and Reeves (1987) suggest the presence of more than just the three Fuller concern dimensions of self, task, and impact represented in Fuller's theory of teacher development.

The purpose of this study was to identify the dimensions of concern measured by the TCCL based on a sample of inservice teachers and to examine the validity and reliability of the dimension scores. Three research questions were addressed:

1. What dimensions (factors) of concern are measured by the TCCL?
2. What is the level of test-retest and internal consistency reliability of resulting factor scores?
3. What is the degree of the relationship between scores derived from the dimensions of concern and measures of teacher stress and quality of teacher work life?

Douglas W. Schipull is now Assistant Professor of Education at Concordia University. Carolyn K. Reeves is Professor of Curriculum and Instruction at the University of Southern Mississippi. Richard Kazelskis is Professor of Educational Leadership and Research at the University of Southern Mississippi. Correspondence and requests for reprints should be addressed to Dr. Carolyn K. Reeves, Department of Curriculum and Instruction, University of Southern Mississippi, S. S. Box 5057, Hattiesburg, MS 39406-5057.

Method

Subjects

Subjects for the study were 568 certified teachers employed by a school district in northwest Florida. The mean age of the sample was 41.87 years, with the age range being 21-65 years. The sample was predominately female (84.33%) with an average of 14.71 years of experience. The distribution of the sample by teaching level was as follows: preschool 1.15%, primary (K-2) 19.59%, elementary (3-5) 34.10%, middle school (6-8) 16.36%, and high school (9-12) 28.80%.

Instruments

The Teacher Concerns Checklist (TCCL), Form B, is a 56-item checklist which was constructed by Frances Fuller and Archie George to explore teachers' concerns at different points in their careers. The TCCL items sample a wide range of concerns, including items which deal with routine tasks as well as items which deal with teachers' perceptions of their impact on students. The directions to respondents indicate that some level of concern is present "if you often think about it and would like to do something about it" (George, 1978, p. 33); in this context, "it" refers to the item/area of concern. A Likert scale is used to rate each item. Teachers are instructed to respond to each of the items using a scale from 1 (not concerned) to 5 (extremely concerned).

Reliability and validity studies of the TCCL have not been reported. While some reliability and validity information was reported by George (1978) for the shortened form, the 15-item Teacher Concerns Questionnaire (TCQ), the reliability and validity of the 56-item TCCL have not been documented.

The Quality of Teacher Work Life Survey (QTWLS) is a 36-item survey which is designed to measure teacher satisfaction and stress (Pelsma, Richard, Harrington, & Burry, 1989). For each item, teachers are instructed to respond to two Likert scales. One scale represents degree of satisfaction, ranging from 1 (very satisfied) to 5 (very dissatisfied). The other scale represents degree of stress, ranging from 1 (no stress) to 5 (extreme stress). Although 10 subarea scores can be obtained from the QTWLS, only the stress and satisfaction scores were used in the analyses reported here.

Using Cronbach's coefficient alpha as a measure of the internal consistency reliability, Pelsma et al. (1989) reported alpha values of .89 and .92 for the satisfaction and stress scales, respectively. Test-retest reliability coefficients for both satisfaction and stress were 0.65 and 0.43, respectively, over a one-year time interval. The majority of the validity correlations between the Maslach Burnout Inventory and QTWLS were found to be moderate and significant in the anticipated direction (Pelsma et al., 1989).

The Teacher Stress Inventory (TSI) is a 36-item questionnaire developed by Schutz and Long (1988), representing a revised version of Pettegrew and Wolf's (1982) original 46-item TSI. The TSI is composed of the following seven subareas/factors: Role Ambiguity (RA), Role Stress (RS), Organizational Management (OM), Job Satisfaction (JS), Life Satisfaction (LS), Task Stress (TS), and Supervisory Support (SS). Teachers are instructed to respond to the TSI using a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree." To obtain a score for the TSI, the teacher responses are grouped by subareas and a total score for each subarea is calculated by summing the responses.

Using Cronbach's alpha, Schutz and Long reported the internal consistency coefficients for the 36-item TSI, which ranged from 0.74 to 0.87 for the seven subareas. The 36-item, seven-factor TSI seems to be at least as reliable as the original 46-item TSI and possesses a superior factor structure. Validity was examined by conducting three MANOVAs on the seven factors/subareas. All seven factors discriminated among the levels of the independent variable ($p < .001$). The results were viewed as being supportive of the construct validity for the 36-item TSI (Schutz & Long, 1988).

Procedures

Fifty-nine school principals in northwest Florida granted permission to collect data in their schools. Two sets of packets were assembled. Packet A contained copies of the Teacher Concerns Checklist (TCCL), the Teacher Stress Inventory (TSI), the Quality of Teacher Work Life Survey (QTWLS), a demographic information sheet, and a cover letter to teachers containing instructions and assurance of the confidentiality of their responses. Packet B was constructed to obtain test-retest data for the TCCL and contained the TCCL and a letter explaining the purpose of the second administration of the TCCL. The letter instructed teachers to respond a second time to the TCCL, as follows: "It is very important that you respond sincerely to the items. It is not necessary to try to remember how you responded to each item on the TCCL (two weeks ago) -- just respond with the level of concern you feel now."

Results

Factor Analysis

Item responses to the TCCL were factor analyzed using alpha factor analysis (Gorsuch, 1983; Kaiser & Caffrey, 1965). Alpha factor analysis was employed because it was the analytic technique used in the original factor analysis by George (1978), and because it maximizes the alpha reliabilities of the extracted factors, i.e., maximizes the psychometric generalizability of the factors (Kaiser & Caffrey, 1965), a particularly compelling goal when factor analyzing test items. Squared multiple-correlations were used for initial communality estimates. Ten dimensions were found with eigenvalues greater than 1.0. These 10 factors were rotated orthogonally using varimax. Those items loading highest on each factor and the associated factor loadings are presented in Table 1.

Table 1
Factor Loadings of the TCCL Items

Dimension/Item	Loading
I - CONCERN ABOUT IMPACT	
Whether each student is getting what he needs	.76
Increasing students' feelings of accomplishment	.75
Guiding students toward intellectual and emotional growth	.71
Whether students are learning what they should	.71
Whether students can apply what they learn	.71
Insuring that students grasp subject matter fundamentals	.70
Instilling worthwhile concepts and values	.66
Recognizing the social and emotional needs of students	.65
Helping students to value learning	.62
Slow progress of certain students	.57
II - CONCERN ABOUT STUDENT ACCEPTANCE	
Whether the students really like me or not	.76
How students feel about me	.76
Acceptance as a friend by students	.74
Where I stand as a teacher	.50
Being asked personal questions by students	.47
III - CONCERN ABOUT PROFESSIONAL FREEDOM	
Teaching required content to students with varied backgrounds	.57
The mandated curriculum is not appropriate for all students	.50
Lack of academic freedom	.42
Standards and regulations set for teachers	.33
IV - CONCERN ABOUT TASK	
Too many noninstructional duties	.66
Feeling under pressure too much of the time	.60
The routine and inflexibility of the teaching situation	.45
Working with too many students each day	.44
V - CONCERN ABOUT STUDENT PROBLEMS	
Student use of drugs	.66
Chronic absence and dropping out of students	.54
The values and attitudes of the current generation	.41
VI - CONCERN ABOUT CLASSROOM BEHAVIOR	
Students who disrupt class	.62
Maintaining the appropriate degree of class control	.59
Lack of respect of some students	.59
VII - CONCERN ABOUT SELF	
Doing well when a supervisor is present	.52
Feeling more adequate as a teacher	.45
Meeting the needs of different kinds of students	.41
Being accepted and respected by professional persons	.41
VIII - CONCERN ABOUT PROFESSIONAL ABILITY	
My ability to present ideas to the class	.50
Increasing my proficiency in content	.50
IX - CONCERN ABOUT SCHOOL CLIMATE	
Clarifying the limits of my authority and responsibility	.47
The psychological climate of the school	.45
Understanding the philosophy of the school	.38
X - CONCERN ABOUT INSTRUCTIONAL MATERIALS	
The nature and quality of instructional materials	.61
Lack of instructional materials	.56

Forty of the 56 items received salient loadings on at least one of the 10 factors. Labels were assigned to the 10 factors based on the theme that the items seemed to reflect. Factor I was labeled *Concern about Impact* because the items composing this factor deal with a teacher's concern about his/her impact on students' overall learning and growth (such as, whether each student is getting what he needs, increasing students' feelings of accomplishment, guiding students toward intellectual and emotional growth, whether students are learning what they should, whether students can apply what they learn). Factor II was labeled *Concern about Student Acceptance* because the items composing this factor indicate concern about personal acceptance by students (such as, whether the students really like me or not, how students feel about me, acceptance as a friend by students). Factor III was labeled *Concern about Professional Freedom* because the items composing this factor express concern about the conflict between mandated regulations and professional freedom (such as, teaching required content to students with varied backgrounds, the mandated curriculum not being appropriate for all students, lack of academic freedom). Factor IV was labeled *Concern about Task* because the items composing this factor express concern about the tasks associated with teaching (such as, too many non-instructional duties, feeling under pressure too much of the time, the routine and inflexibility of the teaching situation). Factor V was labeled *Concern about Student Problems* because the items composing this factor deal with concern about the problems and attitudes of students (such as, student use of drugs, chronic absence and dropping out of students, the values and attitudes of the current generation). Factor VI was labeled *Concern about Classroom Behavior* because the items composing this factor indicate concerns about the classroom behavior of students (such as, students who disrupt class, maintaining the appropriate degree of class control, and the lack of respect of some students). Factor VII was labeled *Concern about Self* because the items composing this factor express concern about professional self-esteem (such as, doing well when a supervisor is present, feeling more adequate as a teacher, meeting the needs of different kinds of students). Factor VIII was labeled *Concern about Professional Ability* because the items composing this factor deal with concern about professional skills (such as, my ability to present ideas to the class, increasing my proficiency in content). Factor IX was labeled *Concern about School Climate* because the items composing this factor deal with concern about the school's atmosphere and philosophy (such as, clarifying the limits of my authority and responsibility, the psychological climate of the school, understanding the philosophy of the school). Factor X was labeled *Concern about Instructional Materials* because the items composing this factor express concern about instructional materials (such as, the nature and quality of instructional materials, lack of instructional materials).

Reliability

Both test-retest and internal consistency coefficients were obtained for scores on the TCCL factors. The test-retest coefficients with a two week-time interval between testings ranged from .69 to .77 (Table 2). Cronbach's alpha coefficients derived from the second administration of the TCCL ranged from .71 to .94.

Correlations with Job Stress and Satisfaction

To control Type I error rates, only those correlations between the unit weighted factor scores for the 10 concern factors and subarea scores on the QTWLS and TSI which were significant at the .01 level were considered for interpretation. All significant correlations were positive and low to moderate in magnitude (Table 2). Scores on all concern dimensions except for Concern about Impact were significantly related to stress scores on at least one stress measure. The highest correlations were found between Concern about Task scores and Task Stress from the TSI (.59) and the total stress measure from the QTWLS (.54). Concern about Task scores were positively related to the total satisfaction scores of the QTWLS (.39). Also, scores on Concern about Instructional Materials were positively related to the total satisfaction scores, though the correlation was minimal (.21).

To further investigate the extent to which the concern scores could predict teacher stress and teacher satisfaction, canonical correlation analyses of the relationship between the TSI and concerns scores and between the QTWLS scores and the concerns scores were carried out. Each of the first two possible canonical correlations between the concerns and TSI subtest scores were found to be significant ($Rc = .670, p < .001$ and $Rc = .433, p < .003$). For the first significant canonical relationship, correlations between the measures and canonical variables suggested each of the TSI variables, except Supervisory Support, to be important contributors to the significant relationship, while concern about Task, Classroom Behavior, Professional Freedom, Self, and School Climate were the major contributors from the concern variables. For the second canonical relationship, the major contributors from the TSI variables were Supervisory Support, Role Ambiguity, and Organizational Management, while only the measure of concern about School Climate received a notable loading. Only the first of the two possible canonical correlations between the QTWLS and concerns dimensions was significant ($Rc = .655, p < .001$). Examination of the correlations between each of the measures and the canonical variate indicated that all variables contributed to the canonical relationship.

Table 2
TCCL Factor Score Reliabilities and Intercorrelations with QTWLS and TSI Scores

	Reliability		QTWLS		TSI						
Factor	Test-Retest	Alpha	Satis-faction	Stress	Role Ambig.	Role Stress	Org. Mgt.	Job Satis.	Life Satis.	Task Stress	Super. Support
Impact	.77	.95	.07	.19	.01	.02	.01	.04	.01	.16	-.08
Std. Accept.	.74	.80	.08	.22[a]	.10	.08	.06	.15	.10	.35[a]	.00
Prof. Freed.	.69	.77	.19	.30[a]	.21[a]	.19	.08	.16	.10	.35[a]	.00
Task	.76	.71	.39[a]	.54[a]	.24[a]	.51[a]	.17	.32[a]	.24[a]	.59[a]	.02
Std. Probs.	.73	.80	.12	.24[a]	-.01	.05	-.08	.03	-.06	.18	-.17
Class. behav.	.77	.80	.14	.21[a]	.13	.20[a]	.13	.29[a]	.18	.47[a]	-.01
Self	.74	.80	.06	.30[a]	.15	.11	.07	.12	.21[a]	.40[a]	-.05
Prof. Abil.	.72	.83	-.02	.14	-.02	-.06	.05	.07	.05	.20[a]	-.09
School Clim.	.71	.85	.15	.25[a]	.20[a]	.16	.14	.23[a]	.05	.34[a]	.09
Inst. Mat.	.72	.66	.21[a]	.16	.11	.09	.09	.11	.07	.30[a]	-.03

[a] $p < .01$

Additionally, multiple regression analyses were carried predicting each of the TSI measures and the Stress and Satisfaction measures from the QTWLS using the TCCL factor scores as predictors. The regression equations were designed to determine the unique effect of each of the TCCL factors by first constructing full models and then removing the predictor of interest from the full model and observing the change in squared multiple correlation. To control for Type I errors across the nine regression equations, the .006 level of significance was used based on the Bonferroni method. The results of these analyses are presented in Table 3. Significant multiple correlations were found for each of the stress and satisfaction measures ($p < .001$) with multiple correlations ranging from .374 for Organizational Management to .640 for Stress as measured by the QTWLS. Thus, from 14 to 41% of the variability in the various stress and satisfaction scores could be accounted for with the concerns scores. The most frequently occurring significant predictors of scores on the QTWLS and TSI were concerns about task, classroom behavior, and school climate.

Table 3
Summary of Regression Analyses

	Criterion								
	Role Ambiguity	Role Stress	Organiz. Management	Job Satisfaction	Life Satisfaction	Task Stress	Supervisory Support	Stress (QTWLS)	Satisfaction (QTWLS)
Predictor	R^2_{chg}	R^2_{chg}	R^2_{chg}	R^2_{chg}	R^2_{chg}	R^2_{chg}	R^2_{chg}	R^2_{chg}	R^2_{chg}
Impact	.021^{a}	.000	.005	.005	.005	.010	.000	.006	.000
Stud. Accept.	.002	.003	.036^{b}	.013	.001	.010	.028^{a}	.019	.011
Prof. Freed.	.026^{a}	.003	.000	.001	.010	.003	.001	.007	.001
Task	.019^{a}	.137^{c}	.059^{c}	.087^{c}	.047^{c}	.138^{c}	.016	.121^{c}	.218^{c}
Std. Probs.	.028^{a}	.008	.017	.002	.021^{a}	.004	.014	.001	.011
Class Behav.	.000	.015	.001	.025^{a}	.027^{a}	.032^{c}	.004	.024^{a}	.034^{b}
Self	.000	.017^{a}	.002	.008	.000	.016^{a}	.004	.001	.010
Prof. Abil.	.001	.000	.000	.002	.005	.000	.001	.003	.005
School Clim.	.035^{b}	.002	.028^{a}	.000	.003	.003	.106^{c}	.000	.000
Inst. Mat	.001	.003	.000	.005	.010	.000	.008	.010	.011
Mult. R	.403^{c}	.476^{c}	.374^{c}	.467^{c}	.416^{c}	.640^{c}	.389^{c}	.624^{c}	.636

$^{a}p < .05$
$^{b}p < .01$
$^{c}p < .001$

Discussion

The results of the factor analysis of the TCCL revealed 10 dimensions (factors) of teacher concern which were labeled Concern about Impact, Student Acceptance, Professional Freedom, Task, Student Problems, Classroom Behavior, Self, Professional Ability, School Climate, and Instructional Materials. Eight of the 10 factors are similar to dimensions reported by George (1978), and 7 of the 10 are similar to factors found by Kazelskis and Reeves (1987) using a sample of preservice teachers. The dimensions of Concern about Professional Freedom and Concern about School Climate were not found by either George or Kazelskis and Reeves, and the dimension of Concern about Student Problems was not found by Kazelskis and Reeves. It was noted that the three factors (i.e., self, task, and impact) which provide the basis for Fuller's theory are among the 10 factors, but the items composing the three factors differ to some extent from those items which represent the three factors on the Teacher Concerns Questionnaire (TCQ) developed by Fuller and George (see George, 1978). For example, 9 of the 15 items which compose the TCQ are included in the self, task, and impact factors which emerged in this analysis. However, only two of the five items which comprise the impact dimension of the TCQ (i.e., whether each student is getting what he needs; guiding students toward intellectual and emotional growth) are among the items composing the impact factor in this study, suggesting that the impact dimension of teacher concerns may not be fully defined in the TCQ. It appears that the five items contained in the impact dimension of the TCQ do not adequately reflect impact concerns of teachers and that a better measure of impact concerns may be obtained by using the 10 items that compose the impact factor in this analysis.

It was not surprising that there was less than perfect agreement between the factors identified in the present study and those found by George (1978) or by Kazelskis and Reeves (1987). George's sample included not only inservice teachers but preservice teachers and school administrators as well. Since the Fuller model of teacher concerns suggests that differences in concern are expected between preservice and inservice teachers, the inclusion in George's sample of both groups along with the school administrators may have produced spuriousness in the item intercorrelations, which would have resulted in factors that would not be found in a more homogeneous group. Although the present sample was heterogeneous relative to age, gender, years of experience, and teaching level, it was selected to represent the typical diversity found within samples of inservice teachers. Kazelskis and Reeves (1987) utilized only preservice teachers in their study and found that their factor analysis resulted in a different set of factors.

Test-retest correlations and alpha coefficients obtained for the 10 dimensions indicate that scores for the dimensions are reasonably stable across time and that each is measured with a relatively homogeneous set of items. These results certainly suggest that the 10 dimensions can be measured with a degree of reliability acceptable for research purposes.

The results indicated that scores on each of the 10 concerns dimensions were related to some aspect of teacher stress and job satisfaction. The multiple regression results, however, particularly highlighted the importance of concerns about the task of teaching, classroom behavior, and school climate as important predictors of teacher stress and satisfaction.

The zero-order intercorrelations between scores on the 10 concerns dimensions and scores from the QTWLS and TSI indicated only low to moderate relationships between scores on the teacher concern dimensions and measures of teacher stress and satisfaction. However, as a set, the concern dimensions represent relatively good predictors of both stress and job satisfaction. Since both stress and satisfaction are perceived to be the result of multiple causes (Albertson & Kagan, 1987; Harris, Halpin & Halpin, 1985; Pelsma et al., 1989; Sharp & Forman, 1985) the use of multiple predictors would be most appropriate and most descriptive.

George (1978) was concerned primarily with an investigation of Fuller's hierarchial model of teacher concerns (Fuller, 1969, 1971), which emphasized concern about impact, self, and task; therefore he did not explore further the other concern dimensions which were found in his factor analysis. The additional dimensions of concern identified here, however, appear to offer a means for studying teacher concerns in a much broader context than that outlined by Fuller. Teacher educators and school administrators who are exploring ways to improve teacher effectiveness are likely to find that preservice and inservice training programs are more beneficial if constructed to address the teaching concerns of those receiving the training.

Collectively, the results of the findings reported here indicate that teacher concerns are multidimensional, arising from sources that include personal, contextual, and philosophical issues. Further study is needed to ultimately determine the value of the concern dimensions reported here; however, the dimensions do appear to tap important aspects of the teaching milieu which are related to both teacher stress and teacher job satisfaction.

References

Albertson, L. M., & Kagan, D. M. (1987). Occupational stress among teachers. *Journal of Research and Development in Education, 21*, 69-77.

Fuller, F. F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207-226.

Fuller, F. F. (1971). Relevance for teacher education: A teacher concern model (UTR&D no. 2313). Austin: University of Texas Research and Development Center for Teacher Education.

George, A. A. (1978). *Measuring self, task, and impact: A manual for use of the Teacher Concerns Questionnaire*. Austin: University of Texas Research and Development Center.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Harris, K. R., Halpin, G., & Halpin, G. (1985). Teacher characteristics and stress. *Journal of Educational Research, 78*, 346-350.

Kaiser, H. F., & Caffrey, T. (1965). Alpha factor analysis. *Psychometrika, 30*, 1-14.

Kazelskis, R., & Reeves, C. K. (1987). Concern dimensions of preservice teachers. *Educational Research Quarterly, 11*, 45-52.

Kyriacou, C. (1987). Teacher stress and burnout: An international review. *Educational Research, 29*(2), 146-152.

Pelsma, D. M., Richard, G. V., Harrington, R. G., & Burry, J. M. (1989). The Quality of Teacher Work Life Survey: A measure of teacher stress and job satisfaction. *Measurement and Evaluation in Counseling and Development, 21*, 165-176.

Pettegrew, L. S., & Wolf, G. E. (1982). Validating measures of teacher stress. *American Educational Research Journal, 19*, 373-396.

Raschke, D. B., Dedrick, C. V., Strathe, M. I., & Hawkes, R. R. (1985). Teacher stress: The elementary teacher's perspective. *The Elementary School Journal, 85*, 559-564.

Schutz, R. W., & Long, B. C. (1988). Confirmatory factor analysis, validation and revision of a teacher stress inventory. *Educational and Psychological Measurement, 48*, 497-511.

Sharp, J. J., & Forman, S. G. (1985). A comparison of two approaches to anxiety management for teachers. *Behavior Therapy, 16*, 370-383.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 35-40

Preservice Teachers' Views on Standardized Testing Practices

Neelam Kher-Durlabhji, Lorna J. Lacina-Gifford, Richard B. Carter, and Randall Jones
Northwestern State University

This study determined preservice teachers' views of high-stakes testing. Two spring and two fall cohorts (n =128 and 140, respectively) of preservice teachers were asked to rate 17 score enhancement strategies for likelihood of use and appropriateness. Results indicated that there is a positive correlation between strategies likely to be used and strategies considered appropriate. Findings of this study suggest that preservice teachers can make acceptable judgments about appropriateness/inappropriateness where the extremes of the continuum of ethical/unethical score enhancement strategies are concerned but fail to do so in the intermediate range of the continuum. Either the preservice teachers have not been exposed to the entire continuum of ethical/unethical test preparation activities in the course of their training or their responses are tempered by the "reality of high-stakes testing."

Background

Like the gambler who bets the limit on a full house, education today is betting higher on student outcomes than ever before. The game is the high-stakes world of testing. This game was implemented in the hopes of improving public education (Shepard, 1992). Herman, Golan, and Dreyfus (1990) define high-stakes testing as situations where teachers, schools, and districts are rated or ranked based upon achievement test scores. Madaus (1988) adds that high-stakes testing may include tests used "for the certification of teachers, promotion of students from one grade to the next, award of a high school diploma, assignment of students to remedial classes, allocation of funds to a school or district, award of merit pay to teachers on the basis of their students' test performance, certification or re-certification of a school district, and placement of a school system into 'educational receivership'" (p. 30).

While some researchers support the application of "measurement driven instruction" (Millman, 1981; Popham, 1981; Popham, Cruse, Ranking, Sandifer, & Williams, 1985), most reviews cite the negative effects associated with high stakes testing. Anderson (1992) notes that, while such endeavors may seem worthwhile on the surface, there is evidence that teachers may be narrowing the focus of their instruction to those areas that are covered in the test. Since standardized tests cover only a narrow range of educational objectives, this is a questionable practice. Anderson concludes that it may even be causing teachers to inadvertently teach a section of the test. This evidence is supported by the findings of Madaus (1988).

Potter and Wall (1992) indicate that one outcome of the education reforms of the 80s has been to create a high stakes testing environment. In their report on the effects of the South Carolina Education Improvement Act, they noted that the increased emphasis on testing, especially high school exit exams, had some negative effects on the dropout rate and in overall grade retention. This effect was most pronounced in some demographic groups, especially nonwhite males.

A strong objection to the practice of testing is made by Glasser (1990), who views standardized testing as limiting the quality of education delivered to the pupils. He believes that quality education cannot be measured by multiple choice items, and chides American education for coercing students into accepting substandard education for the sake of test scores.

The authors of *A Nation at Risk* (National Commission on Excellence in Education, 1983) point out that "minimum" standards on competency tests had become the "maximum," i.e., acceptable. Shepard (1992) suggests that nearly a decade after *A Nation at Risk*, standardized tests have severely limited what students are learning. In Glasser's (1990) view, competency standards were entirely too low, and education was accepting low quality work.

Neelam Kher-Durlabhji is an Assistant Professor in Educational Psychology at Northwestern State University in Natchitoches, LA and Lorna J. Lacina-Gifford is an Associate Professor of Education at Northwestern State University, Natchitoches, LA. Richard B. Carter is an Assistant Professor in Educational Psychology now at Texas Tech University, Lubbock, TX. Randall Jones is a Graduate Assistant at Northwestern State University, Natchitoches, LA. Correspondence regarding this paper should be addressed to: Dr. Neelam Kher-Durlabhji, Division of Education, Northwestern State University, Natchitoches, LA 71497, FAX (318) 357-5092.

While high-stakes testing has been demonstrated to affect the behavior of teachers, students, parents, and administrators, little research has been conducted on its effects on student teachers. The purpose of this study is to determine the views of student teachers toward test preparation practices used in schools.

This research is descriptive and correlational rather than experimental and therefore does not explicitly test any theory. It is an attempt to determine if there is a match between preservice teachers' views of appropriate score enhancement strategies (test preparation practices) and the views espoused by "experts" in the field. Discussions concerning the ethics of various score enhancement strategies used by teachers are mentioned frequently in core educational psychology courses and upper level courses in tests and measurement.

Preservice teachers are exposed to the views of measurement experts such as Haladyna, Nolen, and Haas (1991), Madaus (1988), Mehrens and Kaminski (1989), and Mehrens, Cole, and Popham (1991). These discussions also include a recognition of the discrepancy between opinions articulated by the "experts" and teachers' actual actions when faced with an impending standardized test. Thus issues related to "measurement driven instruction" are systematically addressed in teacher preparation programs.

Madaus (1988) notes six principles that describe the consequences of measurement driven instruction and its effects on teacher and pupil behavior and the test itself. He has aptly captured the flavor of the discussions that occur in our tests and measurement courses. To paraphrase Madaus, the six principles are as follows: First, it does not matter whether or not the results of a test are important; if the students, teachers, and administrators perceive them to be important, they are. Principle 2 states that the more a quantitative indicator like a test is used to monitor social processes, the more it will distort those processes. Third, if important decisions are made based on test results, then teachers will teach to the test. Next, high-stakes tests will eventually define the curriculum. Fifth, the form and format of the questions (i.e., true-false, multiple choice, etc.) eventually defines the curriculum. Finally, because test results have become indicators of future educational life, test results will become a major goal of schooling, rather than a useful but fallible indicator of achievement. Madaus concludes his review by stating that "the irony is that the use of tests as the principal measure of worth destroys these tests' ability to serve as an accurate indicator of student attainment" (p. 36).

Most of the empirical research in "high-stakes" testing has been conducted with teachers, principals, and school supervisors or superintendents. Preservice teachers' views have been markedly absent. As teacher educators charged with the responsibility of preparing future teachers, we considered it important to determine the views of student teachers who had been exposed to issues related to high-stakes testing. We expect that the findings of this study will provide insight into curriculum development.

With these issues in mind, this study was designed to answer the following questions:

1. What score enhancing strategies would preservice teachers *frequently* use with their students?
2. What score enhancing strategies would preservice teachers consider *appropriate to use* with their students?
3. Is there a relationship between score enhancing strategies that preservice teachers consider appropriate and the frequency with which they would use the same strategies in the classroom?
4. Are there differences in the patterns of responses of the fall and spring cohorts of preservice teachers?

Method

Participants

Data in the form of responses to two close-ended questionnaires were procured from two fall and two spring cohorts of preservice teachers. The fall cohorts were comprised of 74 and 66 respondents, and the spring cohorts, 47 and 91 respondents, respectively. Of the total number of respondents, 81% were female and 19% were male. The average age of the respondents was 23.5 years and they ranged in age from 22 to 50 years. The ethnic makeup of the respondents was 78% white, 13% African-American, 2% Native Americans, 1% Hispanic, and 1% Asians. Five percent of the respondents did not respond to the question on ethnic background. All were enrolled in the undergraduate teacher preparation program at a small southern public university. The questionnaires were completed by each group in a group setting at the end of their student teaching semester.

Instrumentation

The various score enhancement strategies rated by the preservice teachers for likelihood of use and appropriateness of use were generated from a review of the literature. Specifically, the theoretical framework articulated by Haladyna et al. (1991), Mehrens and Kaminski (1989), and Mehrens et al. (1991) were used as a guide in developing the survey items. Haladyna et al. (1991) discussed at length various test preparation activities used by teachers. The authors discuss the impact of these activities on "test score pollution" (p. 4) and present these activities on a continuum based on the degree of ethicality.[1] Mehrens et al. (1991) also identified instructional strategies used in preparation for high stakes achievement tests and debated the defensibility of the various strategies. The items generated for the present study were directly based on the test preparation strategies articulated by the above mentioned authors. The survey instrument consisted of two similar questionnaires. Each questionnaire contained 17 items reflecting teacher strategies for test score enhancement (see Appendix A). In the first questionnaire, preservice teachers were asked to rate the 17 items for *frequency of use* on a 6-point scale. In the second questionnaire, the same 17 items were rated for *appropriateness of use*. The 17 items in each questionnaire were randomly ordered to minimize order effects.

The *frequency of use* questionnaire was administered first, followed by the *appropriateness of use* questionnaire. Since both questionnaires contained the same items, responses on the first questionnaire were likely to have a carryover effect on the responses to the second questionnaire. The researchers believed that the carry over effect would be minimized if the teachers were asked to indicate frequency of use in the first questionnaire and appropriateness in the second one.

Results

In general, the pattern of responses of the four cohorts were quite similar; hence a decision was made to collapse the two spring and fall cohorts. Thus the descriptive data presented here are based on a sample of 268 participants.

Responses to the questionnaires were originally categorized along a 6-point scale. In the analysis, the scale points were collapsed to form a trichotomy. Thus responses were categorized as "agree," "neutral," or "disagree" with the presented score enhancement strategies, along the two dimensions of *appropriateness* or *likelihood of use*.

Likelihood of Use

The percentage of respondents who were likely to use the strategies or considered them appropriate is shown in Table 1. A majority of preservice teachers are likely to instruct their students in test taking strategies, encourage students to do their best on standardized tests, send notes home to the parents to elicit their support in motivating students, check students' completed answer sheets to ensure proper marking, develop a curriculum based on standardized test objectives, teach to the test objectives, and present alternate forms of the test or other commercially prepared score boosting activities for practice. Nearly all the preservice teachers indicated that they would not dismiss low achieving students from test taking, change students' completed answer sheets, give hints or clues during the test, allow more than the allocated time for test taking, or present verbatim items from the test for practice. Nearly two-thirds of the respondents disagreed with the strategy of no special test preparation.

Table 1
A comparison of likelihood of use and appropriateness of various test score enhancement strategies

	%	%	Correlation	
Strategies	Likely	Appropriate	Spring[1]	Fall[2]
1. Encourage students to do their best	95.9	96.3	.24	.31
2. Teach test taking skills	84.0	92.9	.45	.32
3. Check student's completed answer sheets	80.7	78.4	.70	.80
4. Send note home to parents to elicit cooperation	69.5	75.1	.66	.58
5. Practice alternative form of test	52.4	54.3	.61	.49
6. Use commercial materials	46.1	57.2	.63	.51
7. Teaches objectives based on standardized test	46.1	33.8	.48	.49
8. Teach according to test objectives	39.4	36.4	.50	.50
9. Develop curriculum based on test content	32.0	36.8	.61	.47
10. Rephrase wording of questions	20.4	13.8	.59	.65
11. Present actual test items for practice	11.5	9.7	.53	.54
12. No special preparation	7.1	4.5	.58	.45
13. Allow more time for allocated time for test taking	3.3	1.9	.71	.37
14. Give hints or clues	3.0	2.2	.81	.37
15. Change completed answer sheets	2.6	1.9	.70	.52
16. Change answers of low achieving students	1.1	0.4	.28	.67
17. Dismiss low achieving students from test taking	0.0	0.0	.27	-.02

*Note: All correlations except the one marked with * are significant at $p < .01$

[1] $n = 128$

[2] $n = 140$

Appropriateness

The pattern of responses to the "appropriateness of use" survey was remarkably similar to the "likelihood of use" survey. All of the 17 correlations were positive and significant ($p < .01$). These date are presented in Table 1.

Cohort Comparisons

In general, the pattern of responses of the two cohorts were similar. Chi-square tests were conducted to determine if the responses of the fall and spring cohort were independent. Of the 17 score enhancement strategies rated for likelihood of use, only one strategy was statistically significant ($p < .05$) and two others were marginally significant ($p < .10$). A similar result was obtained from chi-square tests on the 17 score enhancement strategies rated for appropriateness. Thus it can be surmised that the cohorts were more alike than different in their views regarding the likelihood of use and appropriateness of the various score enhancement strategies presented to them.

Discussion

Haladyna et al. (1991), Madaus (1988), Mehrens and Kaminski (1989), Mehrens et al. (1991), and Shepard (1992) have appraised the ethics of a variety of test preparation activities. In their view, training in testwiseness skills, checking answer sheets, and increasing students' motivation through appeals to students and parents are "ethical" test preparation activities, whereas developing curriculum based on the test, preparing objectives based on test items, or using items similar to those on the test or commercially prepared score boosting activities are "unethical." Presenting items verbatim from the test or dismissing low-achieving students on testing day are considered "highly unethical" test preparation activities. The ethicality of the 17 score enhancement strategies used in this study was rated based on these studies, and is presented in Table 2.

Table 2
Ethicality of Score Enhancement Strategies Based on Criteria Articulated by Haladyna, Nolen, and Haas (1991) and Mehrens and Kaminski (1989)

Degree of Ethicality	Score Enhancement Strategies
Unethical	1. Prepare teaching objectives based on items on the standardized test.
Unethical	2. Teach according to the test objectives.
Highly unethical	3. Alter (change) the completed answer sheets of your students.
Highly unethical	4. Rephrase the wording of items for students having difficulty with the test.
Highly unethical	5. Give hints and clues to students as they take the test.
Highly unethical	6. Dismiss low achieving students from taking the test.
Highly unethical	7. Present items from the test for practice.
Ethical	8. Carry on regular activities with no special preparation for the test.
Unethical	9. Develop curriculum based on content of the standardized test.
Ethical	10. Encourage students to do their best on the test.
Unethical	11. Give the students an alternative form of the test for practice.
Highly unethical	12. Allow students more time for the test than allocated.
Unethical	13. Use commercial materials specifically designed to improve test performance.
Ethical	14. Check your student's completed answer sheet to see if it is properly filled out.
Ethical	15. Send a note home to parents to help prepare child for test.
Ethical	16. Teach test taking skills.
Highly unethical	17. Change the answers of low-achieving students.

Results of this study reveal that preservice teachers' views are congruent with the views of "experts" for strategies listed as "ethical" or "highly unethical" by them. However, the test preparation strategies considered "unethical" by Haldayna et al. are considered "appropriate" by the preservice teachers and are likely to be used by them. Thus, the data suggest that preservice teachers can make acceptable judgments about appropriateness or inappropriateness when the extremes of the continuum of ethical-unethical test preparation activities are considered, but fail to do so in the intermediate range of the continuum. Likely explanations for this result are either the preservice teachers have not been exposed to the entire continuum of ethical/unethical test preparation activities in the course of their training, or their responses are tempered by the "reality of high-stakes testing."

In the broader context of the role of standardized testing in education, this research suggests the need for further study, in particular about the role of the teacher in preparing students for these tests. In the more specific context of the ethics of score enhancement strategies used by teachers, our study reveals a gap between the views of educational researchers and the views of preservice teachers. Previous studies have already documented that educators vary considerably on opinions regarding test preparation and administration practices they see as "cheating." It is likely that teachers consider any practice that boosts test scores to be legitimate (Haladyna et al., 1991). There is urgent need to incorporate a systematic discussion of issues related to appropriate test preparation practices in teacher education programs and at the inservice level.

References

Anderson, J. (1992, April). *Using the norm references model to evaluate Chapter 1*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 350 315)

Glasser, W. (1990). *Quality schools*. New York: Hawthorne.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Herman, J., Golan, S., & Dreyfus, J. (1990, November). [Untitled paper.] Paper presented at the annual meeting of the California Educational Research Association, Santa Barbara, CA.

Madaus, G. F. (1988). The distortion of teaching and testing: High stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46.

Mehrens, W. A., Cole, N. S., & Popham, W. J. (1991, March). *Defensible/indefensible instructional preparation for high stakes achievement tests: An exploratory trialogue*. Paper presented at the joint annual meeting of the American Educational Research Association and the National Council for Measurement in Education, Chicago, IL.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practices, 8*, 14-22.

Millman, J. (1981). Protesting the detesting of PRO testing. *NCME Measurement in Education, 12*, 1-6.

National Commission on Excellence in Education. (1983). *A nation at risk*. Washington, DC: Author.

Popham, W. J. (1981). The case for minimum competency testing. *Phi Delta Kappan, 63*, 89-92.

Popham, W. J., Cruise, K. L., Ranking, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement driven instruction: It's on the road. *Phi Delta Kappan, 66*, 628-635.

Potter, D. C., & Wall, M. E. (1992, April). *Higher standards for grade promotion and graduation: Unintended effects of reform*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 348 750)

Shepard, L. A. (1992). *Will national tests improve student learning?* (Report No. CSE-TR-342). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 348 382)

Footnotes

[1] The authors recognize there might be some disagreement among the scholarly community about discussing the ethics of various test preparation strategies along a continuum. However, the cited authors do place the strategies along an ethical continuum and discuss them in terms of appropriateness of use.

Appendix A: Score Enhancement Strategies

1. Prepare teaching objectives based on items on the standardized test.
2. Teach according to the test objectives.
3. Alter (change) the completed answer sheets of your students.
4. Rephrase the wording of items for students having difficulty with the test.
5. Give hints and clues to students as they take the test.
6. Dismiss low achieving students from taking the test.
7. Present items from the test for practice.
8. Carry on regular activities with no special preparation for the test.
9. Develop curriculum based on content of the standardized test.
10. Encourage students to do their best on the test.
11. Give the students an alternative form of the test for practice.
12. Allow students more time for the test than allocated.
13. Use commercial materials specifically designed to improve test performance.
14. Check your student's completed answer sheet to see if it is properly filled out.
15. Send a note home to parents to help prepare child for test.
16. Teach test taking skills.
17. Change the answers of low-achieving students.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 41-54

Lessons in the Field: Context and the Professional Development of University Participants in an Urban School Placement

Janet C. Richards
The University of Southern Mississippi/Gulf Park

Joan P. Gipe
University of New Orleans, Lakefront

Ramona C. Moore
University of New Orleans, Lakefront

Field experiences are now an integral part of most teacher education programs. But research that looks at university/K-12 connections generally has excluded descriptions of program characteristics, ignored the specific context in which these initiatives take place, and disregarded their impact on university participants (Carter, 1990; Knowles, Cole, & Presswood, 1994; Moore, 1994; Zeichner, 1984; Zeichner, Tabachnick, & Densmore, 1987). Consequently, the knowledge base concerning field experiences remains weak and inconclusive (Lanier & Little, 1986). The qualitative inquiry reported here (a) describes the particular sociocultural aspects of an urban elementary school serving as the context for an award-winning reading/language arts early field experience and (b) takes a close look at how the sociocultural factors of this urban elementary school contribute to the professional development of preservice teachers and university supervisors who work within that context.

Theoretical Perspective:

A fundamental factor affecting what preservice teachers learn in their field placement may be the school context in which teaching occurs (Gipe, Duffy, & Richards, 1989; Richards, Moore, & Gipe, 1994). For example, when preservice teachers are placed in K-12 teaching contexts they have opportunities to (a) become aware of their students' needs (e.g., "These kids are just like kids everywhere who need love and acceptance"); (b) adopt more constructivist views and practices (e.g., "Kids learn best when they can discover and explore"); and (c) develop a good understanding of content-specific knowledge (e.g., "Those word identification lessons worked. The kids are starting to skip unknown words and think about what words might make sense instead of just stopping to sound them out"). On the other hand, the contextual aspects of field placements may influence preservice teachers to (a) become preoccupied with group management concerns (e.g., "I'm not going to let them say one word out of turn!!"); (b) come to consider students with different values, customs, and language as adversaries (e.g., "I can't understand some of them and some give me trouble"); (c) develop more custodial, impersonal, rigid views about teaching (e.g., "I had to give one boy a warning because he gave me two sentences instead of one"); (d) hold fast to previously-acquired "teacher-as-information-giver" beliefs (e.g., "It doesn't work when I let them speak out. Nobody learns anything"); and (e) concentrate solely on survival needs rather than self-reflection (e.g., "I would like to send ten kids to the principal--at least then I could get something done!!").

The contextual characteristics of field placements have the capacity to impact university supervisors' professional development as well.[1] For example, supervisors in charge of a university/school connection have opportunities to learn about the conditions under

---

An earlier paper presented at the World Congress of the International Reading Association, Buenos Aires, Argentina, July 1994. Janet C. Richards is on the faculty of the Division of Education and Psychology at The University of Southern Mississippi/Gulf Park. Joan P. Gipe is a faculty member in the Department of Curriculum and Instruction at the University of New Orleans, Lakefront. Ramona C. Moore is a doctoral candidate at the same institution. Please address correspondence regarding the article to Dr. Janet C. Richards, Division of Education and Psychology, The University of Southern Mississippi/Gulf Park, 730 East Beach Boulevard, Long Beach, MS 39560.

[1] For this inquiry, professional development is defined as changes over time in teachers' professional knowledge base, attitudes, beliefs, and practices (Burden, 1986; Kagan, 1993).

which some teachers must teach and some students must learn (Weiner, 1993). Action research projects and fruitful collaboration with seasoned classroom teachers are also possibilities. On the minus side, university supervisors may become frustrated and overextended because (a) their work in K-12 schools is not supported by university or school system administrators; (b) they must spend considerable time mentoring and soothing anxious preservice teachers before, during, and after class; and (c) they feel isolated and apart from their colleagues and the mainstream of their university. As a result, university supervisors may come to believe that their time spent out in a school setting is not worth the effort expended. Clearly, unless the contextual factors of a "host" school are considered and the effects of those factors on university participants are taken into account, such initiatives may not achieve the more positive outcomes possible (Copeland, 1981; Feiman-Nemser & Buchman, 1987; Guyton & McIntyre, 1990; Zeichner, 1990). "But we currently know very little about these context-specific effects" (Zeichner, Tabachnick, & Densmore, 1987, p. 27). Although dimensions of a field placement have the capacity to impact university participants' professional development, little attention has been given to the school contexts in which preservice teachers and their supervisors work (Knowles, Cole, & Presswood, 1994; Zeichner, 1986).

This inquiry, told through the combined "voices" of three university supervisors who also served as participant researchers, supplies explicit information about the sociocultural factors of an urban elementary school and discusses how those contextual aspects impact university participants in both negative and positive ways. The researchers acknowledge a vested interest in the program. However, despite their personal involvement, the credibility of their efforts was established through structural corroboration (i.e., "the use of multiple sources and types of data to support . . . [their] . . . interpretation[s]") (Pitman & Maxwell, 1992, p. 748).

The Inquiry

Methodology

Tenets of qualitative inquiry guided our research. Qualitative methods are especially appropriate when researchers wish to provide "rich, descriptive data about the contexts, activities, and beliefs of participants in educational settings" (Goetz & LeCompte, 1984, p. 17). We designed the inquiry using triangulation since "the act of bringing more than one source of data to bear on a single point . . . [and] designing a study in which multiple cases, multiple informants, or more than one data gathering are used can greatly strengthen the study's usefulness for other settings" (Marshall & Rossman, 1989, p. 146). We collected data throughout each fall and spring semester for 3 years. Data sources were field notes of formal and informal observations; interviews; information conversations; and artifacts--texts, which themselves are implicated "in the everyday construction of social reality" (Atkinson, 1990, p. 178). The artifacts were preservice teachers' dialogue journal entries; pre- and post-semester written metaphors about teaching; interpretations of a series of seven researcher-devised reading/language arts illustrations that served as a projective technique (e.g., see Harmin & Gregory, 1974 and The Thematic Apperception Test, 1943); and post-semester reflective statements. (Refer to Appendices A through D for examples of these artifacts.)

At the end of each semester we collated all of the data sets for each study participant (i.e., notes documenting observations, interviews, and conversations; dialogue journals; metaphors; interpretations of the reading/language arts illustrations; and final reflective statements). Using constant comparative methods (Glaser & Strauss, 1967), we then scanned, compared, and cross-checked the aggregated data in order to identify common themes or patterns in our own and in the preservice teachers' professional development (Atkinson, 1990; Borko, Lalik, & Tomshin, 1987; Goetz & LeCompte, 1984). In addition, we examined the preservice teachers' dispositions toward reflective thinking and their acquisition of content-specific knowledge (i.e., understanding of current reading/language arts theories and corresponding instructional practices). In the second stage of analysis we engaged in round table discussions focused on our interpretations of the data. Through this interactive process of dialoguing and revisiting the data in an inductive manner, we came to a consensus about recurring themes and patterns. The following themes emerged: (a) the school context in which the program operates is extremely difficult; (b) most of the preservice teachers entered the field experience with feelings of anxiety and apprehension; (c) over half of the preservice teachers entered the program believing that good teaching is the transmission of knowledge and few developed reflective tendencies; (d) over half of the preservice teachers became preoccupied with group management concerns; (e) over the course of each semester the majority of the preservice teachers came to value the field experience, became more aware of their students' needs, developed confidence in their abilities to teach multicultural, urban students, and gradually constructed a good understanding of current literacy theories and practices; and (f) the university supervisors benefited considerably from the experience.

Program Schedule, Orientation, and Activities

Every Monday and Wednesday morning during the fall and spring semesters, the preservice teachers attend classes at Bayview Elementary School (a pseudonym). For the first hour (8:00 - 9:00), they learn about current theories and practices pertinent to teaching language arts effectively to all students, including linguistically different students. From 9:00 to 10:00 they become teachers, struggling and learning on-the-job how to put theory into action and assuming responsibility for implementing reading/language arts strategies in a "real world" situation. At 10:00 they participate in another hour of lectures and seminar discussions. Topics include the reading process, performance-based assessment, and word identification and comprehension strategies.

The program is guided by a constructivist view of learning. For example, discussion sessions with preservice teachers focus on issues such as how human beings learn best (i.e., when they can explore, discover, reason, and continuously interact with their environment) and the benefits of giving students some responsibility for their own learning (Harste, Short, & Burke, 1988; Vygotsky, 1986). The program also emphasizes the importance of teachers reflecting about their work in order to solve educational problems in a thoughtful, deliberate manner (Dewey, 1933; Grossman, 1992). For instance, seminar discussions center on topics such as why all third graders in some school systems must receive reading instruction from third grade basal readers or who ultimately is responsible if a student receives overly-harsh punishment from a teacher.

Because of our beliefs about the benefits of holistic literacy instruction, a literature-based curriculum guides our work with the elementary students. Typical instructional sessions include elementary students and their preservice teachers (a) reading, talking, and writing about literature selections; (b) planning, writing, or editing stories; (c) corresponding in dialogue journals with one another; (d) interacting in visual and performing arts activities; (e) participating in literacy learning games created by the preservice teachers; and (f) engaging in reading comprehension and writing strategies.[2]

The Elementary School Context: We selected Bayview Elementary School as the context for our program because of its idealistic, permissive, student-centered philosophy, urban setting, and culturally diverse, academically at-risk student population. We wanted a teaching context that would introduce the preservice teachers to a view of schools different from their own experiences. In all likelihood, future "teachers who are white or middle class will probably not teach students like themselves" (Grant, cited in Weiner, 1993, p. 110). Therefore, we wanted to prepare our preservice teachers for possible future employment. In this context, we also hoped to be able to challenge their beliefs and previously-acquired conceptions about teaching and help them recognize and perhaps broaden their perspectives. Studies suggest that when preservice teachers are confronted with teaching practices, values, and beliefs which "differ from their own . . . [they are] more likely to examine and reconstruct their own beliefs" (Kagan, 1992, p. 157). Another consideration was the receptiveness of the teachers and principal, Dr. Rob (a pseudonym). Further, unlike the prevailing climate in some urban schools, Dr. Rob and his teachers work to make their school a democratic community by "withstanding institutional pressures for uniform instruction and custodial treatment of students" (Weiner, 1993, p. 121). Therefore, at Bayview we knew that we would feel comfortable and have the freedom to structure our program as we wished.

The school is located in an old red brick, non-air conditioned building. In the spring and fall, temperatures in individual classrooms may reach over 100 degrees. Recently, the school board decided to close Bayview because of safety hazards. For example, the roof leaked considerably, and portions of the ceiling occasionally fell, narrowly missing students. For reasons unknown to us, the school remained open and some minor repairs have now been accomplished. However, the halls and classrooms are dark and dingy; electric light bulbs hang suspended from frayed cords; window shades are torn or missing; and the walls are cracked and peeling. Apparently, the school board does not consider Bayview when allocating money for structural and aesthetic city-wide school improvements.

There is a permissive atmosphere at Bayview School. Students address teachers by their first names, and it is common for students to walk out of their classrooms without asking permission in order to use the bathroom or water fountain or to speak to Dr. Rob concerning their problems with teachers and peers. Students are encouraged to interact and verbalize with one another whenever they wish. Consequently, the noise level throughout the school is high. There is no dress code.

---

[2] In June, 1991, this program was awarded the American Association of Higher Education's Presidents' Award for "Exemplary Work in Accelerating Minority Students' Achievements."

Some teachers wear cut-off jeans and t-shirts, and students dress as they choose.

Bayview is not a large school. There are 12 teachers; one section of each grade level for kindergarten through Grade 5 and two sections of each grade level for Grades 6 through 8. Class sizes vary from 20 in kindergarten to 30 in each of the seventh and eighth grades.

The Elementary Students: Of the approximately 350 students at Bayview, 80% are African-American; 16% Caucasian; and 4% Hispanic-American. Little parental support is offered to the majority of the students. Each school year at least 92% of the students receive government-subsidized breakfast and lunch; most live in nearby low income housing.

The younger students at Bayview particularly enjoy our program. They work in small groups of approximately 10 to 12 students with our preservice teachers, and they relish the extra attention and the holistic, literature-based lessons. However, because of underlying familial and environmental problems, a few students exhibit developmentally inappropriate social conduct, such as biting others when angry; hitting and yelling to alleviate frustrations and to get their own way; and playing and running through classrooms, hallways, and the basement during instructional sessions.

Each semester at least 75% of the students in the upper grades are over age because of on-going academic problems throughout their school careers. Some were "dropouts" for a year or more prior to attending Bayview. "Unable to achieve in school, these . . . [students] . . . see academic success as unattainable and so they protect themselves by deciding school is unimportant" (Comer, 1988, p. 6). Unfortunately, verbal disruptions among the students occur often, and each semester one or two students are suspended or expelled for aggressive behavior, carrying concealed weapons, or selling or using drugs.

The older students bring a laizzes faire attitude with them as they work with the preservice teachers. Consequently, many are uncooperative and participate minimally in literacy lessons. Low self-esteem, a sense of inadequacy, inner conflicts, peer pressures, and chronic anger may contribute to some students deliberately trying to offend the preservice teachers (e.g., "I do not like white people!"). Yet, as the semester continues, many of these same students volunteer to carry the preservice teachers' books and teaching supplies to their cars. At the end of each term some students also express regret that the preservice teachers have completed their work at Bayview.

Most of the students are in need of rich literacy experiences. Data collected over the past 3 years indicate that they have become more motivated to read and write. They also write more and take more risks with their writing. However, the students' oral and written language and reading abilities continue to be under developed for their ages and grade levels.

The Elementary Teachers: Almost half of the teachers at Bayview are recent liberal arts graduates. They are members of a special teaching corps (Teach for America) who volunteer to work for 2 years in schools throughout the United States as they complete courses toward certification. Other teachers have been at Bayview for over 15 years.

Dr. Rob encourages his teachers to design their own curriculum and to teach lessons as they think best. Their individual instructional orientations range from a skills-based "teacher-as-information-giver" focus to a holistic, constructivist "teacher as facilitator" view. Because of Dr. Rob's strongly articulated, idealistic, and "free" student-centered philosophy and the prevailing values of the school, new teachers quickly adopt permissive attitudes toward their students. It is also possible that most of these new teachers enter "teaching to change education--and--society. [Therefore] they [have] a considerable investment in making sure [that] their classrooms [are] democratic or free" (Weiner, 1993, p. 119).

The classroom teachers are supportive of our program. They welcome the preservice teachers wholeheartedly and look forward to having them work with their students. However, despite our efforts as supervisors to involve the classroom teachers in program planning and implementation, they remain minimally involved. Very few adopt any of these lessons they observe. The teachers are very busy and there is little opportunity for them to begin to understand the literacy theories that undergird our program and the preservice teachers' practices.

The Impact of Bayview's Sociocultural Factors on University Participants

Preservice Teachers: Most preservice teachers enter our program with feelings of apprehension and anxiety. Their dialogue journals reflect their fears and lack of self-confidence.[3] For example, John writes: "January 31st . . . Nothing in the world could have prepared me for what I saw today; I am unprepared. They are trying to give me a hard time." Sally's journal mirrors similar concerns:

---

[3] The preservice teachers whose journal excerpts appear in this manuscript have graciously given permission for their writing to be shared. The names of preservice teachers and elementary students are pseudonyms.

"January 31st . . . Somehow I am afraid. I am very vulnerable. I woke up every hour on the hour all night. This school is all run down. Couldn't you find a better school?" The preservice teachers' concerns are legitimate. Like most elementary education majors, the majority "are white, middle-class, and female" (McDiarmid, 1990, p. 12); few have worked with groups of students prior to this experience; most have never closely interacted with African-Americans.

After their initial sessions at Bayview the preservice teachers' frustrations and feelings of estrangement rise. Sample journal entries include, "What am I supposed to do with these kids?"; "Why don't you tell us what to do?"; "We need more time on campus"; "You are just throwing things at us"; and "Many of the students could not understand the words I was saying in the spelling test even though I spoke slowly and distinctly."

The preservice teachers do settle in over time and become more aware of their students' needs (e.g., "These kids aren't as bad as I thought. In fact, they're just like kids everywhere who need love and acceptance"). However, as described in the literature (e.g., see Hoy & Woolfolk, 1990; McNeely & Mertz, 1990), many preservice teachers become preoccupied with group management concerns (e.g., "Next week I'm going to try a behavior management chart. I'm not going to let them say one word out of turn!"). Others come to view their students as adversaries (e.g., "It was a good day. All the kids behaved and Nicky, the one who gives me the most trouble, was absent. Thank God!" and "These kids have no respect for us. One of my students walked in Monday and called me a dog. You can't work with kids like that! I never talked to my teachers that way!").

Most of the preservice teachers enter our program believing that good teaching is the transmission of knowledge (McDiarmid, 1990). As the semester progresses, only a few adopt more constructivist perspectives. Some studies show that school placement or program philosophy do not affect preservice teachers' prior beliefs (Lortie, 1975; McDiarmid, 1990, cited in Kagan, 1992). But according to Feiman-Nemser and Buchman (1987), we may not be doing enough to urge the preservice teachers to examine their previously acquired beliefs about teaching and learning. On the other hand, the teaching placement is difficult. Despite the constructivist emphasis in our program and our own constructivist practices, we know that the preservice teachers have little time to consider alternative philosophies of teaching.

Our program is designed to promote reflective thinking. Therefore, we urge the preservice teachers to think carefully about their work (e.g., "Reflect! Reflect! Tell me why you think things went so well"). We also articulate our own reflective orientations and try to operate as reflective practitioners. Yet, we know that most preservice teachers enter and exit our program displaying minimal or nonexistent reflective tendencies. A few preservice teachers are "natural born" reflectors (e.g., "I've always reflected since I was a little girl. I ask myself 'why' about everything. I LOVE this!"). But, many confuse thinking reflectively with stating procedural facts about their lessons (e.g., "MY REFLECTIONS! I put the kids in a circle. I handed out the papers. Things went okay. That's about it"). Perhaps we expect too much. As one preservice teacher explained to us, "We do reflect. But, we're just learning this stuff. It takes awhile for it to sink in." It also is understandable that the contextual factors of the field placement may influence the preservice teachers to consider their own survival needs first rather than their students' needs (Fuller, 1969; Richards & Gipe, 1988).

On a more positive note, the preservice teachers' work with the elementary students and their responses to the illustrated reading/language arts vignettes show that over the course of the semester they gradually construct a good understanding of current literacy theories and instructional practices. Most of the preservice teachers also come to value the field experience and develop confidence about their abilities to teach multicultural, urban students who are academically at-risk (e.g., "I now know more than any other outsider would ever believe. This has changed me forever"; and "When I registered for these classes I had no idea what I was doing. I felt like someone had put me in a blender and pressed the puree button. But, it has gotten easier. I now know that I can handle anything. It does not matter where I will work").

The University Supervisors: We concur with Weiner (1993) that as supervisors of a university/urban field program our job is labor intensive. We spend considerable time solving problems; teaching demonstration lessons; meeting with classroom teachers; traveling to and from the elementary school; and observing, mentoring, and soothing preservice teachers. While the time given to these activities may not be "valued in the reward structure of . . . [teacher education] institutions" (Tafel & Christensen, 1988, p. 4), we have benefitted a great deal from this experience. We continue to maintain energy as program supervisors, and we remain committed to the program. We have close contact with practicing teachers and their students. We know the values and customs of an urban school and understand under what conditions some teachers must teach and some students must learn. Working together at Bayview School has given us many opportunities to collaborate with one another and to forge close collegial, professional, and personal relationships as well.

We present research findings from our studies conducted at Bayview School at national and international conferences and publish reports in scholarly journals. Our research efforts allow us to "give voice to the otherwise silenced voices of [urban classroom teachers and] students" (Weiner, 1993, p. 133). In addition, supervising this initiative forces us to examine our own orientations. We adjust course content and assignments to align our practices with our views, and we continually reflect about our work in order to ensure that we "practice what we preach."

Because our program is literature-based we also have become more knowledgeable about quality multicultural children's literature. Further, working in a literature-based program motivates us to create reading comprehension and writing strategies compatible with holistic literacy instruction and the elementary students' instructional needs. Most importantly, supervising preservice teachers in a multicultural urban setting helps us to know "well the dilemmas of both teacher education and the joys and difficulties of daily teaching" (Cuban, 1993, p. x).

Final Reflections: Engaging in this inquiry has enabled us to address the contextual realities of our program. We cannot deny that Bayview Elementary School's permissive, student-centered philosophy, and the negative attitudes of many of the older students contribute to making this field placement a difficult assignment for our white and middle class preservice teachers. Their dialogue journal statements, informal conversations, and final reflective statements attest to their anxieties, fears, and frustrations. As Zeichner (1986) indicates, the social conditions of a school can present obstacles to preservice teachers' learning. However, we know that "urban teachers of at-risk students require special preparation" (Weiner, 1993, p. 7). We believe that working in an urban elementary school with multicultural students who are academically at risk prepares our preservice teachers for the realities of their future employment conditions.

At the same time, we recognize that we need to work toward making this field placement less difficult and demanding for our preservice teachers. Toward that end, we are considering implementing activities such as presenting reality-based case studies, demonstration lessons, and role playing simulations, plus encouraging observations at the school site prior to each semester's field experience. These activities will provide our preservice teachers with some background knowledge about Bayview School's values, philosophy, and student population and give the preservice teachers a repertoire of strategies for soothing group conflicts and mediating disputes among students. We also are considering restructuring the program so that the preservice teachers work with fewer students. In this way the preservice teachers will "become familiar and comfortable with individual students so that students become real people rather than categories" (Weiner, 1993, pp. 118-119).

As supervisors of a university/urban elementary school initiative we have discovered that collaboration has a synergistic power. We become energized as we work together solving logistical problems and crises associated with the field program. We are learners along with our preservice teachers. We have insider knowledge of an urban elementary school (Cuban, 1993). And, we remain committed to the program.

In the final analysis, we conclude that working in Bayview Elementary School as university supervisors has had a profoundly positive impact on our professional development. The same is not always true for our preservice teachers. At times, the elementary students' behavior influences many of our preservice teachers to believe that they have not accomplished their teaching goals (e.g., "I cannot shake this feeling of failure"). The negative attitudes and inappropriate behavior of some students also influence the preservice teachers to attempt to counteract the permissive attitude of the school by equating good teaching with keeping students quiet and on task (e.g., "Smaller groups are easier to manage, and I just shiver at what it would be like trying to teach these students as a whole class"). It is also understandable that the difficulty of the field placement impedes the preservice teachers' development of reflective practices and constructivist views about the learning process. However, we know that by the end of the semester, they learn a considerable amount about current literacy theories and practices, and they feel comfortable teaching multicultural students in an urban school. We also recognize that placing our preservice teachers in an urban school environment is a necessity. "First-hand experience and theoretical knowledge about urban schools are essential for virtually all [preservice] teachers who will work with poor, minority students in urban schools" (Weiner, 1993, p. 129).

References

Atkinson, P. (1990). *The ethnographic imagination: Textual constructions of reality.* New York: Routledge.

Borko, H., Lalik, R., & Tomshin, E. (1987). Student teachers' understandings of successful and unsuccessful teaching. *Teaching and Teacher Education, 3*, 77-90.

Burden, P. (1986). Teacher development: Implications for teacher education. In J. Raths and L. Katz (Eds.), *Advances in teacher education* (Vol. 2) (pp. 185-219). Norwood, NJ: Ablex.

Carter, K. (1990). Teachers' knowledge and learning to teach. In W. Houston (Ed.), *Handbook of research in teacher education* (pp. 291-310). New York: Macmillan.

Comer, J. (1988, November). Educating poor minority children. *Scientific American, 259*, 2-8.

Copeland, W. (1981). Clinical experiences in the education of teachers. *Journal of Education for Teaching, 7*(1), 3-16.

Cuban, L. (1993). Foreword. In L. Weiner, (Ed.), *Preparing teachers for urban schools: Lessons from thirty years of school reform.* New York: Teachers College Press.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educational process.* Boston: D. C. Heath.

Feiman-Nemser, S., & Buchman, M. (1987). When is student teaching teacher education? *Teacher and Teacher Education, 3*, 255-273.

Fuller, F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207-226.

Gipe, J., Duffy, C., & Richards, J. (1989, Fall). A comparison of two types of early field experiences. *Reading Improvement*, 254-265.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine.

Goetz, J., & LeCompte, M. (1984). *Ethnography and qualitative design in educational research.* New York: Academic Press.

Grant, C. (1989). Urban teachers: Their new colleagues and curriculum. *Phi Delta Kappan, 70*, 764-770.

Grossman, P. (1992). What models matter: An alternate view on professional growth in teaching. *Review of Educational Research, 62*, 171-179.

Guyton, E., & McIntyre, D. (1990). Student teaching and school experiences. In W. Houston (Ed.), *Handbook of research on teacher education* (pp. 514-534). New York: Macmillan.

Harmin, M., & Gregory, T. (1974). *Teaching is . . . Experiences and readings to help you become the kind of teacher you want to become.* Chicago: Science Research Associates.

Harste, J., Short, K., & Burke, C. (1988). *Creating classrooms for authors.* Portsmouth, NJ: Heinemann.

Hoy, W., & Woolfolk, A. (1990). Socialization of student teachers. *American Educational Research Journal, 27*(21), 279-300.

Kagan, D. (1992). Professional growth among preservice and inservice teachers. *Review of Educational Research, 62*, 129-169.

Knowles, G., Cole, A., & Presswood, C. (1994). *Through preservice teachers' eyes: Exploring field experiences through narrative and inquiry.* New York: Macmillan.

Lanier, J., & Little, J. (1986). Researching teacher education. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 527-569). New York: Macmillan.

Lortie, D. (1975). *Schoolteacher: A sociological study.* Chicago: University of Chicago Press.

Marshall, C., & Rossman, G. (1989). *Designing qualitative research.* Newbury Park, CA: Sage.

McDaniel, J. (1990, April). *Close encounters: How do student teachers make sense of the social foundations?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

McDiarmid, G. (1990). Challenging prospective teachers' beliefs during early field experiences: A quixotic undertaking? *Journal of Teacher Education, 41*, 12-20.

McNeely, S., & Mertz, N. (1990, April). *Cognitive constructs of preservice teachers: Research on how student teachers think about teaching.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Moore, R. (1994). *The impact of a portal school project on teacher development.* Unpublished manuscript.

Pitman, M., & Maxwell, J. (1992). Qualitative approaches to evaluation: Models and methods. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 729-770). New York: Academic Press.

Richards, J., & Gipe, J. (1988, April). *Reflective concerns of prospective teachers in an early field placement.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 296 976)

Richards, J., Moore, R., & Gipe, J. (1994, April). *"This school is a terrible place. The kids don't listen": Contextual influences on preservice teachers' professional growth in an early field placement.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Tafel, L., & Christensen, J. (1988). Teacher education in the 1990s: Looking ahead while learning from the past. *Action in Teacher Education, 10*, 1-6.

The Thematic Apperception Test. (1943). Boston: Harvard University Press.

Vygotsky, L. (1986). In R. Reiber and A. Carton (Eds.), *The collected works of L. S. Vygotsky: Vol. a: Problems of general psychology*. New York: Plenum.

Weiner, L. (1993). *Preparing teachers for urban schools: Lessons from thirty years of school reform.* New York: Teachers College, Columbia University.

Zeichner, K. (1984). Preparing reflective teachers: An overview of instructional strategies which have been employed in preservice teacher education. *International Journal of Educational Research, 11*, 565-575.

Zeichner, K. (1986). Individual and institutional influences on the development of teacher perspectives. In J. Rath & L. Katz (Eds.), *Advances in teacher education* (Vol. 2) (pp. 155-163). Norwood, NJ: Ablex.

Zeichner, K. (1990). Changing directions in the practicum: Looking ahead to the 1990s. *Journal of Education for Teaching, 16*(2), 105-132.

Zeichner, K., Tabachnick, R., & Densmore, K. (1987). Individual, institutional, and cultural influences on the development of teachers' craft knowledge. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 21-59). London: Cassell.

Appendix A

Example of a Preservice Teacher's Dialogue Journal

January 31

Nothing in the world could have prepared me for what I saw today. I've never been in a class that is so disorganized and uncontrollable in my life. I realize it is only the first day and it should get better. We administered the vocabulary test first and most of the students seemed unmotivated to do this test at first. Many of them did not have the pens or pencils to even complete the test. Some of them had to use crayons. I was unprepared for this dilemma. After this we attempted to do the spelling test. I passed out a sheet of paper to everyone in class. I gave the instructions and we administered the spelling test. I was surprised that many of the students could not understand the words I was saying, even though I spoke slowly and distinctly. They obviously had a problem understanding white English vernacular or they were trying to give me a hard time. Finally we handed out the personal journals. In all I had a good positive, and enthusiastic attitude from the students.

February 2

We handed out the journals first today. Already the students seemed less enthusiastic about writing in their journal. I can't seem to find anything to motivate these children. Many of the students to go art and special classes; I feel I really can't hold them back from this. I need to find some other way to motivate them. After the journals, we worked on the name tags. Some of the students work diligently toward making their name tags, but most were once again unmotivated to do anything but talk. I handed out 28 gold medal name tags (made from yellow construction paper) and got back only around 18. Ten of these got "lost in the shuffle".

February 7

Today Beth administered the interest inventory. This was only half as successful as it should have been. The students seemed interested in doing the activity, but they were also interested in doing others things at the same time--like talking. I've realized that the students like to answer questions. Now I have to find the right questions to ask. The interest inventory consisted of questions on cards that each student must answer individually. This did not work. So what we did was ask the questions orally and had the students write individual answers down. I got a lot of responses just by walking around and listening to group discussions.

February 9

We introduced the reading response logs today. Barely half of the students responded, if that much. They listened to the story and probably enjoyed it, but when it comes to do writing lessons of any sort whether it's in their journals or doing interest inventories or just a reading response log, they did not seem motivated in any way. We talked about how they grade kids in class today. All these kids see is whether something is for a grade. If it's not it doesn't affect them. They have no motivation to do good in their evaluations at the end of the year turn out. They have no respect for us. They know that our work matters only to us and to you.

February 21

I read a story to them today. I chose a story from an Encyclopedia Brown Book. I read to them "The Case of the Ticking Clock." I really did this to see what kind of listening skills they had---whether they would listen to the story and then possibly, using listening comprehension, solve the story. We took important facts out of the story and then evaluated them to come to a conclusion. This worked with the students; the ones that were interested. I tried to get them to do a reading response log on this story, but it seemed more successful just talking about the story rather than having them write about it.

February 28
Beth taught a lesson on "Kubla Kahn." Many of the students weren't paying attention, even though they were given a copy of the poem. The students had no desire to learn about this poem, much less the desire to write a reading response log about it. I am becoming more and more aggravated with these students everyday. It seems that I'm not able to do hardly anything with these kids. I can't connect with them.

March 2
Today we split up the class and I took my kids outside. That was a big mistake. Four of them hit the gate. I finally got most of them corralled. I had to send one of them up to the office because he wouldn't behave. I would like to send ten of them up there at least then I would be able to get something done. The lesson went well though, after I got them settled. It was kind of un-uniform but I think the kids enjoyed it and learned something.

March 7
Today Beth taught a lesson. I moderated the class. I can tell Beth is losing patience with the students. I don't blame her. It seems every time I write in my log it's complaining about my students. I hardly ever have something good to say about them. Well anyway some of the students completed the story frame. A couple of the students who I least expected to finish, did. I was quite surprised at the participation. After the students wrote in their journal I discussed ideas for the mural. They seemed to like the idea of having a gold medal for the mural.

March 9
We started the mural today. All the students were enthusiastic about completing the mural. I had lots of participation. Everyone was given a blank piece of paper and told to draw a picture pertaining to the Olympics (our classroom theme). I was once again surprised at the participation rate, but was quite disappointed at the end product of some pictures. Some of the pictures had no correlation to the Olympics.

March 14
We were busy finishing up the mural today. Everyone was quite helpful. The students were busy cutting, pasting, sticking on stickers, and drawing more pictures. At times the class got a little hectic because the students got off task. I tried to keep them all on task but that is not always possible. The students were overly proud of the mural they had created.

March 15
Today went fifty percent good, fifty percent bad. Everyone was quiet and on task. Too bad, half of the class was on task playing cards. I tried to get them to put away the cards but they wouldn't. Joan, the teacher, had a talk with them and they continued to play so I assumed that she kind of condoned it or it was a reward for something I did not know about. I let it go. The rest of the class and I had a group conversation on the opening scene in the movie "Menace to Society." It pertained to racism and protection of property. I think the discussion opened some eyes and hopefully changed some of the students attitudes toward interpretation of racism.

March 21
Today we went into the cafeteria and started on ideas for our books. The students had some good ideas about stories. Some had trouble thinking of stories, so I thought I would bring my guitar in and try a lesson with them.

March 23
Today we sang the blues. The students were having trouble thinking of stories and ideas. They wouldn't work because they had "writers block." So I pulled out my guitar and we sang the blues. This was to show them that they could easily think of a story off the top of their head. I provided the riff, they provided the lyrics and story. It worked well.

April 4
Today I tried to get the students to finish their books. All I got was broken promises. I'm working on it. I'm gonna do it. When it was all said and done, I had all the work from only one student.

April 6
It was totally impossible to get the students to work downstairs so instead of giving even more time for their books, I decided to do a word map. The word map was fine. I really got the students interested. It didn't turn out well but I got the job done. The class was quite interested and then right when you (Dr. Richards) walked up, they went nuts. I mean really psycho. They were running around. They were hitting on each other. I was really embarrassed.

April 11
I decided to do my data this week. It started out very poorly, but I implemented a very "neat" discipline procedure. I had two people's names on a referral and if anyone else's name was to go on the referral all three would be sent up to the office. Well this worked great. It kept everybody in the group disciplining and controlling each other and it kept the stragglers from other classes away. I was proud of the lesson and the discipline strategy.

April 13
Today we finished the data by discussing drug and alcohol abuse and their consequences. Also I got the students to do some writing work. They write sequels to the story/song I read to them. I can't believe it.

April 18
I was sick today. Sorry!!!

April 20
Today I did oral retelling taping of the story "Jumanji." I could tell many of the students didn't pay attention, but the last girl, Shondora, was outstanding. You could have sworn that she was reading a summation off a page somewhere. I was so surprised. Today Leslie got bitten by a student and I was in the office when she came in. She was crying hysterically. This upset me. I asked her what was wrong and if I could do anything, she just ignored me. This upset me more. She's obviously still mad at me. Oh well.

April 25
Today we did oral retelling with the book "Nettie's Trip South." The students were not paying attention; I could tell. I even warned the students that they would be retelling the story and to pay attention. They didn't. After, Pete and Andy told me that the students said they couldn't retell the story because they couldn't hear the story. If they would have shut their mouths they could have heard the story.

April 27
Today the students did the sustained silent reading of the book "Wagon Wheels." Many of the students were insulted at the reading level they were asked to read. I told them since this book was so easy they shouldn't have any problems with it. They still were uninterested.

April 29
As suggested by you (Dr. Richards) and my fellow classmates, I took off of work today to come to Bayview School. I took four of my students down to the library to work on their books. I got all the work done for all four students, even though I still had to constantly ride one of them to get her work done. I think you'll be proud of how the books turned out. That's the news and I am outta here.

May 2
We didn't get anything done today because the students were restless in their anticipation for their field trip to the Superdome. We had a lot of flaring tempers between the students and between the students and teachers because many students were upset that they had not brought in their permission slips. Joan said that there was no use in trying to get anything done because of the field trip, but Beth got a small lesson done with some of the students.

May 4
Beth suggested a small party and maybe a video for the students. After Monday I really wouldn't give them anything. They really don't deserve it, but why make the good ones suffer for the bad. So, I sucked up my views and agreed on it. After all they are just kids and it's better that they are here than out on the streets.

Appendix B

Examples of Preservice Teachers' Pre- and Post-Semester Metaphors about Teaching

Stephanie's Pre-Semester Metaphor

The Coach Approach
The teaching and learning process can be considered as coaching a team. The teacher is the head coach and the students make up the team (along with the coach). As in coaching, the teacher explains, demonstrates, and puts in to practice the skills that are to be learned. However, the coach does take into consideration the individual's potential. The coach expects the student to perform to their own potential. This in turn, makes the team up as a whole. The students as a whole make up the learning team. The teacher guides the students like a coach would and as the students practice the skills, they master the skill by working as a team. By helping each other they all become winners!

Stephanie's Post-Semester Metaphor

No Strikes Out!
Teaching is like a baseball team cooperating, and the whole team, including the coach, giving it their all. The teacher is just that, a coach, giving it more than 100%. The coach is also a friend with high standards for herself and the team. In order to make it to the World Series they all, as a team, have to work together, cooperate, and respect one another.

Leslie's Pre-Semester Metaphor

A Salesman

I see a teacher's role as being that of a facilitator. The metaphor that comes to mind is one of a salesman. You are selling a product (an education) to a prospective buyer (the student). In order for the buyer to want the product, the seller has to make it exciting and viable. The seller cannot get bogged down using technical terms or the buyer will quickly lose interest. The buyer needs to try out the product and use it according to his or her need. The use can be refined later on as needs dictate. A buyer has many different needs and it is up to the seller to fill these needs.

Leslie's Post-Semester Metaphor

The Car Salesman

The car salesman is me, the teacher, while the customers are my students. The salesman has to be prepared and have extensive knowledge of his inventory to inform the customer. He also has to listen to the customer in order to find out what he or she needs. The salesman has to be enthusiastic and motivating when making a sales pitch to get the customer's interest. The salesman has to have a fair pricing scheme so the customer won't feel ripped off. The pricing scheme fits in with the grading scale and whether a student feels it is fair. If a sale is made, the salesman/teacher has been successful with this sales pitch and the customer/student is completely satisfied.

Elizabeth's Pre-Semester Metaphor

The Canoe Trip Approach

Teaching is like taking a canoe trip. The principal takes you upstream and lets the teacher, canoe, and her children (the canoers) off. The canoers and canoe are sent down stream to continue on their adventure. Sometimes the trip is easy and the current brings you downstream like when the children may catch on to what is being taught. Sometimes when the river is high the canoers have to paddle long and hard like when the children have a hard time catching on to a lesson. The river may even be so rough in spots that the canoers flip over in their canoes. Troubled kids need to be pulled out of the water, their canoes bailed out, and again sent on their way. When the canoe trip is over, the children will be so excited that they will be ready for another adventure. Some will get used to falling in the water, but they'll learn not to panic and to finally pull themselves out.

Elizabeth's Post-Semester Metaphor

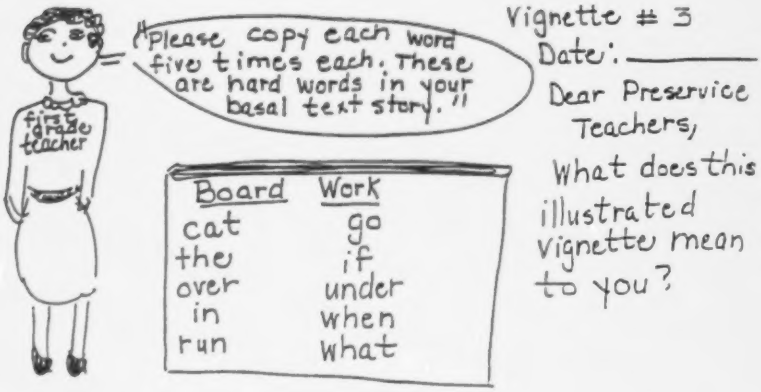
White Water Rafting

I think teaching is more like white river rafting than just a canoe trip. The principal brings us up river and helps us (the students and I) get our boat in the water. As a teacher, I sit in the back and watch over the riders (the students). The water is rough in spots and I give directions as I see they are needed. The students take turns guiding the boat. If a rider falls overboard, I'll fish them out and set them going again. If one of my students is failing, I'll do my best to make sure I push the student to do their best and pass. When the water is smooth, the riders are able to navigate to places of their choice. I keep them on the main route, but I'll let them explore.

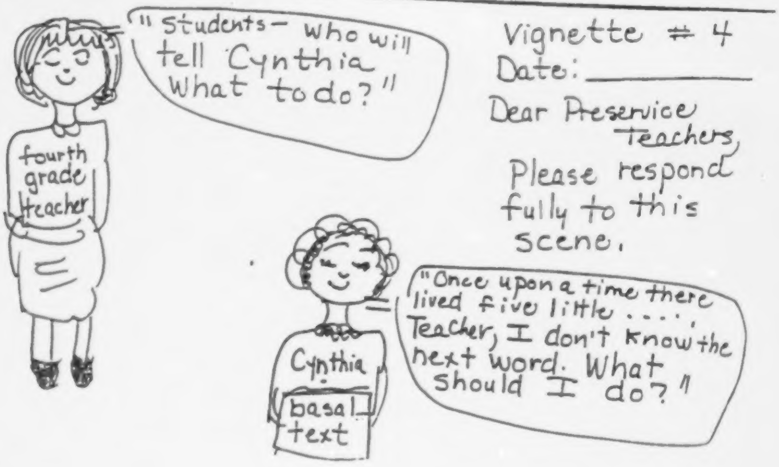
White river rafting is a faster pace than canoeing. Like teaching is exciting, sometimes its slow and sometimes it goes by so quick. Once you get started there is no getting off. The river is also ever changing as the water breaks down and builds up rocks. Teachers have to be ever changing. They have to keep abreast of the new information and get rid of extra baggage.

Appendix C

Examples of the Researcher-Devised Reading/Language Arts Illustrations


"Please copy each word five times each. These are hard words in your basal text story."
first grade teacher
Board Work
cat go
the if
over under
in when
run what

Vignette # 3
Date: ________
Dear Preservice Teachers,
What does this illustrated vignette mean to you?


"Students - who will tell Cynthia what to do?"
fourth grade teacher
"Once upon a time there lived five little .... Teacher, I don't know the next word. What should I do?"
Cynthia
basal text

Vignette # 4
Date: ________
Dear Preservice Teachers,
Please respond fully to this scene.

BEST COPY AVAILABLE

Appendix D

Example of a Preservice Teacher's Post-Semester Reflective Statement

Overall, I would say my entire time at Bayview School was successful in helping me to grow professionally. I did not enjoy many of the times that I had to teach, but I am grateful for the experience that they have given me. It would not be an understatement if I had problems with the group of students that I had to teach. The reasons for this are many and are discussed throughout the rest of this reflection.

In the area of literacy instruction I feel more confident about my ability to go out and teach this subject. I do not claim to be all-knowing in this subject area, for this is my first time in taking instruction. Being all-knowing is not important, what is important is that I now have an idea of what literacy instruction is about and also where to look for ideas for lessons. As to my earlier beliefs about literacy instruction, I do not think that I had many other than that I wished to try to instill in my prospective students the love that I myself have for reading and that I thought that basal texts were an easy way out for a lazy teacher. I still feel that reading is essential to life as food, water, and air. One of the new or changed ideas I have from attending this block is that I now see that the basal texts are what you make of them. I see that they are very comprehensive, and that they can be an integral beginning to any curricula. I have also seen many good strategies that I had never seen when I was a student. Things like Reader's Theater and language experience stories let me see that literacy instruction does not have to be just boring drills or mundane activities, but can teach as well as be fun.

I think my ideas about how children learn are unchanged. This does not seem to me to be a problem, either in the course or me. I think that over the past year or so that I have begun to really solidify my viewpoint in this area. This viewpoint is that children learn in many different ways. Truly an earth shaking statement this is not, but I truly believe in what it means. In teaching (and also coincidentally in disciplining), I think that a teacher is doing his or her student an injustice if they try to use one method in teaching. I feel that this view is what the literacy block tries to teach my peers and me. This view is further seen in the various activities that we were taught in this class as well. Teaching strategies like language experience stories or open word sorts, two very different activities, allow for the individual performance of every student involved. These ideas show that my beliefs in how children learn are strengthened, not changed because of this class.

My group management skills were tested severely by this teaching experience. To say that I had no control at times over some of my students is a mild claim. I would like to save most of the discussion about this for the next part of the reflection, but I do know one area I learned something. Being a special education major means I have to take behavior management classes, but those did not help me with this class very much (for many reasons, see below). I did learn one thing, sometimes it is good to split a group down into smaller parts even though it may mean more work. Smaller groups are definitely easier to manage and I just shiver at what it would have been like trying to teach these students as a whole class. I do not want you to believe that I think that when I get in the real world I will be able to always have a smaller group. In this situation it was a major and welcomed aid in dealing with these students.

I would say that the greatest area of success was in my own personal growth as not only a person, but as a teacher. I now know that I could teach in an upper elementary environment and not lose my sanity and be committed, or worse, because of the experience. I never thought I would be able to handle that type of environment, but I now know I have both the patience and ability to cope and teach in this type of situation. This is an incredible growth for me and what is so amazing about it all is it stems from some of my most negative experiences at Bayview School. My students were at times unruly and that was on a good day. I now have reflected over all of my experiences and I know there were many factors that led to this negativity. As you already know the teacher is having her own problems with discipline in her class, and has changed plans often. On top of this I enter the picture, teaching for just one hour twice a week, and never ever having had a previous experience in dealing with such discipline problems before. Over the course of the whole experience, I have had both good experiences in teaching as well as totally negative ones. I even think that the totally free nature of the school, especially the third floor, contributes to the problems I, as well as the teacher, had with the students. But still I see success in it all. Where I once thought I could only work with and have patience to deal with small kids, I now know I can deal with bigger ones. I used to joke that I would probably go totally crazy if I had to deal with older kids, but I see that that was a useless exaggeration and a totally inaccurate belief I held about myself.

I am truly grateful for the experience at Bayview School, but I still feel regret. I truly wish I was able to have more success in my teaching experiences with these children. I feel as if I have failed my profession. I know that many of the things that happened were beyond my control, but still I feel grateful. I know I can teach. Some of my lessons at Bayview went well and that was an incredible feeling, but I cannot shake this feeling of failure. At times over the experience I ranged from elation over a good day to feeling like a failure to outright hatred for many of the students. I do not like that last emotion, but I realize that I am only human and can only take so much. Success is that I was patient, understanding, and able to cope with those emotions. I will always feel as if I failed because I could not reach all of my students, but realistically I know that I may never reach every student I teach. I know that I have grown as a person and I did have some successful lessons. I know that I got through this with success because of my own abilities, the help of my peers, and the guidance and encouraging words of my teachers. I pity those who cannot go through this experience.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 55-62

Locus of Control, Social Interdependence, Academic Preparation, Age, Study Time, and the Study Skills of College Students

Craig H. Jones, John R. Slate, and Irmo Marini
Arkansas State University

We investigated the relationship of students' study skills to their locus of control, social interdependence, academic preparation, age, and study time. Participants were 266 students enrolled in seven sections of an introductory psychology course at a university in the mid-South. Study skills were related to locus of control, age, expected course grade, and study time. Social interdependence, high school grades, perceived academic preparation, number of hours enrolled, and number of hours employed were unrelated to study skills. The need to address attitudinal and motivational variables in study skills programs is discussed.

Since the 1970s, study skills training programs have been adopted in over 50% of all 4-year state institutions as a means of decreasing student attrition (Cowart, 1987). These programs often vary substantially in length and topics covered, with some institutions offering study skills training for academic credit and others requiring only at-risk students to enroll. Despite the popularity of such programs, empirical findings are inconsistent as to whether students' academic achievement improves as a result of participating in study skills training (Kirschenbaum & Perri, 1982).

At least four factors may explain why study skills courses are often unsuccessful in enhancing student achievement. First, study skills programs tend to be general in nature rather than focusing on the specific strengths and weaknesses of individual students (Jones, Slate, Mahan, Green, Marini, & DeWater, 1994). This approach undoubtedly makes inefficient use of available time. That is, unfocused programs probably devote a considerable amount of time teaching students skills they have already acquired.

Second, study skills programs generally fail to deal with attitudinal variables that may influence students' study behaviors. Even if students learn more effective study skills, whether or not they utilize these new skills probably depends to a large degree on their perception as to whether use of these skills will make a difference. For example, Cone and Owens (1991) argued that students who possess an internal locus of control may be more likely to implement new study skills than are students with an external locus of control. That is, students who have an external locus of control may learn study skills but not use them because of beliefs that grades are influenced primarily by fate, luck, or other forces beyond their control. Although research on the overall relationship between locus of control and academic achievement has produced inconsistent results (Lefcourt, 1976; Phares, 1976), Munro (1981) reported that students with an internal locus of control persevered in college longer than did students with external locus of control. More recently, Agnew, Slate, Jones, and Agnew (1993) reported that students with an internal locus of control exhibited significantly better study skills than did students with an external locus of control.

Another attitudinal variable that may influence the use of academic skills is social interdependence. This variable reflects the extent to which people are interested in working cooperatively, competitively, or independently in attaining goals. Johnson and Johnson (1975) reviewed the literature on social interdependence and academic achievement and noted that this research has focused on differences between students with competitive and cooperative orientations. Based on this review, Johnson and Johnson concluded that competition was superior to cooperation when tasks were extremely simple (e.g., routine drill) and required little help from others. Cooperation produced greater achievement on more complex tasks such as learning to solve mathematics problems. In addition, students working in cooperative groups outperformed students working competitively on

---

Craig H. Jones and John R. Slate are Professors in the Department of Counselor Education and Psychology at Arkansas State University. Irmo Marini is an Assistant Professor and Coordinator of the Rehabilitation Counseling Program at Arkansas State University. Please address correspondence regarding the paper to Craig H. Jones, Ph.D., Department of Counselor Education and Psychology, Arkansas State University, P.O. Box 940, State University, AR 72467-0940. We would like to express our appreciation to the three reviewers who reviewed an earlier version of this manuscript for their helpful comments.

tasks involving the memorization and retrieval of information such as learning names and dates. Thus, the social interdependence orientations of students should affect their use of study skills in academically important ways.

Third, many study skills programs are voluntary. Schwartz (1992) noted that students who attend voluntary study skills sessions have higher grade point averages (GPAs) than do students who do not attend voluntary programs. As a result, positive effects found for voluntary study skills programs tend to disappear when prior GPA is controlled. This suggests that students with poor academic backgrounds not only have lower levels of academic skills than do other students but are less likely to realize that they are in need of academic help as well.

Fourth, study skills courses may have minimal impact on academic achievement because college students may not devote sufficient time to studying to make effective use of the skills they acquire. Michael (1991) has noted that so many activities compete for college students' time and attention that they are unlikely to study until it is absolutely necessary. Indeed, the results of a number of studies indicate that a high percentage of college students typically wait until the night before a test to study for it (Agnew et al., 1993; Jones, 1989; Jones, Slate, & Kyle, 1992; Jones et al., 1994). Thus, amount of time spent studying, and factors that affect it should be related to the effective use of study skills. Perhaps the two most important factors related to the amount of time students have available are number of credit hours taken and number of hours employed (Astin, 1993).

Before study skills programs can be improved, researchers must improve understanding of the factors that influence student use of the skills that are taught. Thus, we conducted the present research to explore further college students' use of effective academic skills. Because previous research had already identified characteristic study behaviors in this student population (Jones, 1989; Jones et al., 1994), the present study was focused on the relationships between study skills and attitudinal and temporal variables. The specific research questions addressed were: (a) Are college students' study skills related to locus of control? (b) Are college students' study skills related to their orientations toward social interdependence? (c) Are college students' study skills related to the adequacy of their academic preparation and to their expectations for success in a college course? (d) Are college students' study skills related to temporal constraints, or variables such as time spent studying, number of credit hours enrolled, and number of hours spent in outside employment?

Method

Participants

Participants were 266 undergraduate students enrolled in seven sections of an Introduction to Psychology course at a university in the mid-South. Because this course is part of the general education curriculum, students tend to be highly representative of incoming students at this university. With permission of the instructor, students anonymously completed the questionnaire packet during regular class periods.

The sample included 147 women and 119 men. Most students were either freshmen (n = 174) or sophomores (n = 92), with a few juniors (n = 4) and seniors (n = 7). The mean age was 20.6 years (SD = 5.07) with a range from 17 to 52. There were 244 whites, 15 African-Americans, and 7 students of other ethnic backgrounds (e.g., Asian or Hispanic).

Instruments

The questionnaire packet began with a section requesting that students provide basic information on themselves. The demographics measured included age, sex, ethnicity, and class status. Adequacy of preparation was measured by asking students to provide their high school grade point average (HSGPA), and report the extent to which they believed their high school had adequately prepared them for college. Previous research has shown that HSGPA is the single best predictor of college grades (Astin, 1971), and that self-reports of GPAs are highly accurate (Fetters, Stowe, & Owings, 1984). Expected success was measured by asking them to provide their expected course grade. Students were also asked to provide the average amount of time they spent studying each week, number of credit hours for which they were enrolled, and the number of hours they were employed each week. Although college students over-report how much time they spend studying, this over-reporting has the effect of adding a constant (Schuman, Walsh, Olson, & Etheridge, 1985). Thus, self-reports of study time should not be used to estimate actual study time (i.e., an absolute measure), but can be used appropriately to order students according to how much they study (i.e., a relative measure). Self-reports of study time were, therefore, appropriate in the present study because the purpose of this measurement was simply to order students on this variable.

The first instrument to which students responded was the Study Habits Inventory (SHI). The SHI consists of 63 true-false items designed to assess the typical study behaviors of college students (Jones & Slate, 1992). There are 30 items that describe effective study behaviors and 33 items that describe ineffective study behaviors. Items indicating ineffective study behaviors are reverse scored and responses are summed to yield an index of academic skills ranging from 0 to 63, with high scores reflecting better study skills than do lower scores.

Jones and Slate (1992) have reported on the psychometric properties of the SHI. Reliability is good, with a 2-week test-retest reliability coefficient of +.82. Internal consistency is also high, with a coefficient alpha mean across studies of +.85. In the current study, the coefficient alpha was +.86. The SHI has been validated, in part, through correlations with college students' grades, with individual studies yielding *r*s ranging from +.16 to +.54. Validity has also been established by finding predicted correlations between SHI scores and measures of other variables including dualistic thinking ($r = -.33$), procrastination ($r = -.46$), and locus of control ($r = -.62$). These findings indicate that students with high SHI scores are less dualistic in their thinking, procrastinate less, and exhibit more of an internal locus of control than do students with low scores on the SHI.

Next, students completed the Academic Locus of Control Scale for College Students (ALC). The ALC has 28 true-false items related to personal control over academic outcomes (Trice, 1985). Scores range from 1 (strongly internal locus) to 28 (strongly external locus). In this study, the coefficient alpha of the ALC was +.70. This finding is similar to previous studies in which the coefficient ranged from +.68 (Agnew et al., 1993) to +.70 (Trice, 1985).

Finally, students completed the Social Interdependence Scale (SIS) developed by Johnson and Norem-Hebeisen (1979). The original SIS consists of 22 items on a 7-point True-False scale, but the response format was changed to a 5-point Likert-type format (strongly agree, agree, neutral, disagree, strongly disagree) for this study. The SIS has three separate scales, cooperative, competitive, and individualistic, consisting of 7, 8, and 7 items respectively. Scores on the cooperative and individualistic scales could range from 7 to 35, and scores on the competitive scale could range from 7 to 40. The higher the score on each scale, the more cooperative, the more competitive, or the more individualistic subjects consider themselves to be. Scores on these scales are relatively independent so that a student could conceivably receive a high score on all three scales. This separate scoring procedure was employed because research prior to the development of the SIS had not consistently supported the argument that these orientations reflect different points on the same continuum (Johnson & Norem-Hebeisen, 1979).

The validity of the SIS was established previously through a series of factor analytic studies (Johnson & Norem-Hebeisen, 1979). These studies supported the independence of the competitive and cooperative scales. Scores on the individualistic scale were negatively correlated with scores on the cooperation scale and positively correlated with scores on the competitive scale. That is, students with an individualistic orientation tend to be less cooperative and more competitive than are other students. Coefficient alphas reported in the validity studies ranged from .84 to .88. The internal consistencies using the modified response format in this study was +.94 for the cooperative scale, +.85 for the competitive scale, and +.73 for the individualistic scale.

Data Analysis

The overall relationships between study skills and demographic characteristics, academic preparation, temporal factors, and attitudinal variables were investigated using stepwise multiple regression with SHI scores as the criterion variable and the other measures as predictor variables. A forward entry procedure was used with statistical significance at the .05 level as the entry/removal criterion. The categorical variables of sex and perceived adequacy of high school preparation were coded as dummy variables. Expected course grade was coded using the standard equivalents for calculating grade point averages (i.e., A = 4, B = 3, C = 2, and D = 1).

For those predictor variables found to be significantly related to study skills by the regression analysis, the specific nature of these relationships was investigated with discriminant analysis using responses to the individual SHI items as the discriminating variables. A forward stepwise procedure based upon Wilks' lambda was used with statistical significance at the .05 level as the entry criterion. Group classifications for continuous variables were created by dividing the sample into thirds and eliminating the middle third. Although this procedure results in data loss that would not occur with a simple median split, contrasting the upper and lower thirds results in a clear contrast by eliminating misclassifications near the median caused by measurement error (Dane, 1990).

Results

The mean SHI score for students in this study was 32.1 ($SD = 9.2$), indicating that they typically performed only 51% of the behaviors assessed by the SHI appropriately. This mean is comparable to the means of 33.0 and 34.2 found in previous research with students at the same university (Jones, 1989; Jones, Slate, Marini, & DeWater, 1993). The mean ALC total score for the sample was 12.1 (*SD* = 4.3), indicating that most students tended slightly toward an internal locus of control. This is comparable to the mean of 12.5 found previously with students at this university (Agnew et al., 1993). The mean Cooperative score was 22.7 (*SD* = 7.5), the mean Competitive score was 24.4 (*SD* = 6.4), and the mean Individualistic score was 19.9 (*SD* = 4.9). Thus, the students were above the mean on all three scales, but not extremely so. There were 168 students who believed their high school education had adequately prepared them for college, 96 who believed their high school preparation was inadequate, and 2 students who did not respond to this item. Students reported a mean HSGPA of 3.27. They were very optimistic regarding their expected course grade with 138 expecting an A, 90 expecting a B, 33 expecting a C, and only 4 expecting a D, with one student not responding to this item. Students reported studying a mean of 93.9 minutes per day (*SD* = 78.2) with a range from 0 to 510 minutes. Students were enrolled for a mean of 12.9 credit hours (*SD* = 2.35). The mean number of hours employed each week for the entire sample was 14.9 hours (*SD* = 13.8). When the 98 students who reported that they were not employed were excluded, however, the mean number of hours worked rose to 23.6 per week.

Regression Analysis

The regression of SHI scores on the demographic characteristics, academic preparation, temporal factors, and attitudinal variables produced a statistically significant equation, $F(4, 212) = 45.04$, $p < .01$, which accounted for 45% (adjusted R square = .45) of the variance in study skills. The four variables that contributed to this equation are displayed in Table 1. The strongest relationship ($\beta = -.52$) was between study skills and locus of control. Students who had an internal locus of control tended to have better study skills than did students who had an external locus of control. The second strongest relationship ($\beta = .19$) was between study skills and age, with older students tending to have better study skills than did younger students. Expected course grade and study time made roughly equal contributions to the prediction of study skills although in opposite directions. The lower the expected grade, the better study skills students displayed. As expected, study skill use tended to improve as students devoted more time to study.

Table 1
Regression of Study Habits Inventory Scores
on Demographics, Academic Preparation,
Temporal Factors, and Attitudinal Variables.

Variable	B	Standard Error	β	ΔR^2	*t*	*p*
Locus of Control	-1.12	.12	-.52	.38	-9.57	.01
Age	.34	.09	+.19	.05	3.72	.01
Expected Grade	-1.48	.64	-.11	.01	-2.29	.02
Study Time	.01	.01	+.11	.02	2.17	.03
Constant	39.95	2.82				

Note. Variable listed in order of entry. $R^2 = .46$; Adjusted $R^2 = .45$

Discriminant Analyses

Locus of control. Students in the upper third of the ALC scores ($n = 73$, range = 14-22) were contrasted with students in the lower third of the ALC distribution ($n = 93$, range = 2-10). The resulting discriminant function was statistically significant, $\chi^2 = 155.55$ ($df = 21$), $p < .01$, and accounted for 59% of the between groups variance (canonical correlation = .77). The group centroids were -1.13 for students with an external locus of control and .98 for students with an internal locus of control.

Following the recommendation of Tabachnick and Fidell (1983), the SHI items with pooled-within-groups correlations of .30 or greater were used to identify the discriminant function. The 10 items used to identify the discriminant function are listed in Table 2. The positive correlation coefficients indicate that students with an internal locus of control were more likely to respond appropriately to this item than were students with an external locus of control. Thus, these items indicate that locus of control was strongly related to motivational-attentional difference between students. Students with an external locus of control exhibit much greater difficultly getting down to work and maintaining attention to their work than do students with an internal locus of control. Students with an external locus of control are also less likely than are students with an internal locus of control to seek help from instructors when needed.

Age. This analysis contrasted students in the upper third of the age distribution ($n = 89$, range = 20-52 years) with students in the lower third of this distribution ($n = 87$, range = 17-18 years). The discriminant function was statistically significant, $\chi^2 = 98.15$ ($df = 27$), $p < .01$, and accounted for 46% of the between groups variance (canonical correlation = .68). The group centroids were .92 for the younger students and -.90 for the older students.

Table 2
Study Habits Inventory Items with Pooled-Within-Subjects Correlations of .30 or Greater in the Discriminant Analyses

Variable	SHI Item	*r*
Locus of Control		
	I have to wait for the mood to strike me before attempting to study.	+.48
	I have trouble settling down to work and do not begin studying as soon as I sit down.	+.47
	I frequently get up, write notes to friends, or look at other people when I should be studying.	+.41
	I have a tendency to doodle or daydream when I am trying to study.	+.39
	I spend too much time on loafing, movies, dates, and so forth that I should be spending on my coursework.	+.36
	When sitting in my classes, I have a tendency to daydream about other things.	+.36
	I frequently test myself to see if I have learned the material I am studying.	+.33
	When I have difficulty with my work, I do not hesitate to seek help from my instructor.	+.32
	I try to complete assigned readings before my instructor discusses them in class.	+.32
	I try to space my study periods so that I do not become too tired while studying.	+.30
Age		
	I have trouble settling down to work and do not begin studying as soon as I sit down.	-.37
Expected Grade First Function		
	I use the facts learned in school to help me understand events outside of school.	+.41
	I tend to skip over the boxes, tables, and graphs in a reading assignment.	+.39
	I work out personal examples to illustrate general principles or rules that I have learned.	+.36
	I try to think critically about new material and not simply accept everything I read.	+.35
	I use the facts I learned in one course to help me understand the material in another course.	+.32
	I often do not have reports ready on time, or they are done poorly if I am forced to have them in on time.	+.31
Expected Grade Second Function		
	I keep a special indexed notebook or card system for recording new words and their meanings.	+.33
Study Time		
	I spend too much time on loafing, movies, dates, and so forth that I should be spending on my coursework.	+.37
	I read by indirect (diffused) light rather than by direct light.	-.33
	I study most subjects with the idea of remembering the material only until the test is over.	+.33

The SHI item with a pool-within-groups correlation above .30 is listed in Table 2. This item indicated that older students had less difficulty settling down to study than did younger students.

Expected course grade. Because only four students expected a grade of D, this analysis was restricted to three groups, that is, students expecting As, students expecting Bs, and students expecting Cs. The first discriminant function was statistically significant, $\chi^2 = 149.27$ ($df = 52$), and accounted for 57.7% of the explained variance which was 30% of the between groups variance (canonical correlation = .55). The group centroids were .60 for the students expecting As, -.56 for students expecting Bs, and -1.00 for students expecting Cs. Thus, the function most strongly discriminates students expecting As from students expecting Cs. Students expecting Bs fall midway between the other two groups.

The six SHI items with pooled-within-groups correlations above .30 are listed in Table 2. The positive correlations indicate that students expecting As were most likely to report appropriate behavior on these items, and students expecting Cs were least likely to report appropriate behaviors on these items. With one exception, these items reflect a greater emphasis on meaningful learning by students expecting As than by the other students. That is, students expecting As try to use what they learn in school to understand events outside of school, attempt to learn the additional material presented in boxes and tables, develop personal examples to improve their understanding of concepts, think critically about new material, and try to relate material from various courses to improve understanding.

The second discriminant function was also statistically significant, $\chi^2 = 149.27$ ($df = 52$), accounting for 42.3% of the explained variance which was 24% of the between groups variance (canonical correlation = .49). The group centroids were .10 for the students expecting As, -.59 for students expecting Bs, and -1.21 for students expecting Cs. Thus, the function most strongly discriminates students expecting Bs from students expecting Cs. Students expecting As fall midway between the other two groups.

Only one SHI item had a pooled-within-groups correlation above .30. This item (see Table 2) indicated that students expecting to receive Cs were more likely than other students to have a special system for learning new terminology. Students expecting Bs were least likely to employ this study behavior.

Study time. This analysis contrasted students in the upper third of the study time distribution ($n = 86$, range = 120-345 minutes/day) with students in the lower third of this distribution ($n = 75$, range = 0-45 minutes/day). The resulting discriminant function was statistically significant, $\chi^2 = 110.995$ ($df = 24$), $p < .01$, and accounted for 53% of the between groups variance (canonical correlation = .73). The group centroids were .99 for students in the high study time group and -1.13 for students in the low study time group.

The three SHI items with pooled-within-groups correlations of .30 or greater are presented in Table 2. These items indicate that study time is related to motivational variables. Students who invest little time in studying are more likely to want to remember material only until the test is over and are more likely to loaf when they should be studying than are students who invest more time in studying.

Study Time, Employment, and Course Load

Although study time was related to study skills, number of credit hours taken and number of hours worked were not. This was surprising given that course load and work load should affect study time (Astin, 1993; Greenberger & Steinberg, 1986). Thus, the relationship of course load and work load to study time was investigated further by calculating the correlations between these measures. Study time was significantly related to both hours employed, $r(258) = -.15, p < .01$, and credit hours taken, $r(253) = .25, p < .01$. The more hours students worked, the less they studied; the more credit hours they were enrolled for, the more they studied. The proportion of the variance of study time in both cases, however, is very small. That is, hours employed accounted for only 2% of the variation in study time, and hours enrolled accounted for only 6% of the variation in study time.

Discussion

Consistent with previous studies (Agnew et al., 1993; Jones et al., 1993), students in the present study exhibited poor academic skills, performing appropriately only 51% of the academic behaviors assessed on the SHI. These findings support the continuing need for study skills programs. The results of this study, however, also showed that students' use of their academic skills vary as a function of a number of factors that, therefore, must be addressed in the development of study skills training programs.

The most powerful predictor of appropriate study skills use was locus of control. As was found in previous studies by Agnew et al. (1993) and Cone and Owens (1991), students who expressed an internal locus of control reported better study skills than did students who expressed more of an external locus of control. These students differed primarily in terms of both motivation to study and the ability to concentrate on academic work. That is, students with an external locus of control were not only less likely to study than were students with an internal locus of control, but were also more likely to be distracted from studying once they had begun.

Age was also a significant predictor of study skills. Although adult students often express concerns that they lack the academic skills needed to compete successfully against younger students (Schlossberg, Lynch, & Chickering, 1989), older students in the present study displayed better study skills than did younger students. Jones et al. (1994) also found that older students had better study skills than did younger students. Thus, older students may often underestimate their ability to be competitive in relation to younger students. The discriminant analysis indicates that the slight advantage older students have over younger students in the use of study skills is related to older students being more able to settle down and get to work. Thus, study skills programs that focus on older students may need to focus less on time management and motivational factors than is needed in study skills programs that focus mainly on traditionally aged students.

Expected course grade was also related to the use of study skills. Interestingly, the higher the grade a student expected, the lower the quality of his or her study skills. The discriminant analyses revealed that students expecting As place more emphasis on meaningful learning than did other students, whereas students expecting Cs were more likely than were other students to have a system for recording new terminology. Thus, one possible interpretation is that students who expect lower grades may engage in more behaviors that technically qualify as appropriate study behaviors than do students expecting higher grades, but that students expecting lower grades focus too much on rote learning.

An alternative explanation for the inverse relationship between expected grades and study skills arises from the fact that this relationship emerged in the regression equation after the effects of locus of control had been statistically removed. Thus, if locus of control and expected grade are related, the obtained relationship between expected grade and study skills could be an artifact of the stepwise regression procedure. This interpretation is supported by the results of a simple one-way analysis of variance of SHI scores based upon expected grade, $F(3, 251) = 7.89$, 7.89, $p < .01$. TukeyB tests revealed that in this analysis students who expected an A had better study skills ($M = 34.6$) than did students who expected a B ($M = 29.7$) or a C ($M = 28.1$), but not students who expected to make a D ($M = 33.8$).

Study skills were positively correlated with time spent studying. This is not surprising given that effective study requires distributed practice and cannot be performed when students cram for examinations. On the other hand, study skills were not related to academic load or to hours employed--even though both of these variables were related to study time. Hours enrolled and hours employed are only weakly related to study time and, therefore, have little relation to *how* students study.

The discriminant analysis indicated that the relationship between study time and study skills is based upon motivational, not temporal, factors. That is, the students who studied the most were the students who were interested in learning course material in a way that would promote long-term retention. Students who were merely interested in passing the next test were more likely to loaf than to study.

Several of the variables investigated in this study were unrelated to study skills. None of the three social interdependence scales (i.e., cooperative, competitive, or individualistic) was correlated with study skills. In addition, high school preparation, whether measured subjectively or with HSGPA, was unrelated to study skills. Thus, there is no assurance that underprepared students will seek study skills training on their own, or that traditional measures used to assess "at risk" status will correctly identify students in need of such training. Only direct assessment of study skills appears likely to identify accurately students in need of study skills training.

Overall, the results of this study suggest that study skills programs must address attitudinal and motivational issues in addition to providing technical assistance on how to study. That is, even if study skills training programs are successful in teaching relevant skills to students who have an external locus of control, these students are unlikely to put the skills they learn to good use. In addition, students who view education as memorizing material long enough to pass the next test are unlikely to invest sufficient time to employ study skills effectively. Readers, however, should remember that the present data are correlational and that direct causal conclusions are not warranted. Thus, study skills programs should also include a formal evaluation plan to assess the effectiveness of the various components of the program.

References

Agnew, N. C., Slate, J. R., Jones, C. H., & Agnew, D. M. (1993). Academic behaviors of agriculture students as a function of academic achievement, locus of control, and motivation orientation. *NACTA Journal, 37*(2), 24-27.

Astin, A. W. (1971). *Predicting academic performance in college*. New York: Free Press.

Astin, A. W. (1993). *What matters in college?* San Francisco: Jossey-Bass.

Cone, A. L., & Owens, S. K. (1991). Academic and locus of control enhancement in a freshman study skills and college adjustment course. *Psychological Reports, 68*, 1211-1217.

Cowart, S. C. (1987). *What works in retention at state colleges and universities*. Iowa City, IA: The ACT National Center for the Advancement of Educational Practices, The American College Testing Program.

Dane, F. C. (1990). *Research methods*. Pacific Grove, CA: Brooks/Cole.

Fetters, W.B., Stowe, P. S., & Owings, J. A. (1984). *High school and beyond: A national longitudinal study for the 1980s* (Report No. TM 840 743). Washington, DC: National Center for Educational Statistics. (ERIC Document Reproduction Service No. ED 249 292).

Greenberger, E., & Steinberg, L. D. (1986). *When teenagers work: The psychological and social costs of adolescent employment*. New York: Basic Books.

Jones, C. H. (1989, Spring). Improving students' study skills. *Arkansas School Psychology Association Newsletter*, pp. 4-5.

Jones, C. H., & Slate, J. R. (1992). *Technical manual for the Study Habits Inventory*. Unpublished manuscript, Arkansas State University, AR: State University.

Johnson, D. W., & Johnson, R. T. (1975). *Learning together and alone.* Englewood Cliffs, NJ: Prentice-Hall.

Johnson, D. W., & Norem-Hebeisen, A. A. (1979). A measure of cooperative, competitive, and individualistic attitudes. *The Journal of Social Psychology, 109*, 253-261.

Jones, C. H., Slate, J. R., & Kyle, A. (1992). Study skills of teacher education students. *The Teacher Educator, 28*, 7-15.

Jones, C. H., Slate, J. R., Mahan, K. D., Green, A. E, Marini, I., & DeWater, B. K. (1994). Study skills of college students as a function of gender, age, class, and grade point average. *Louisiana Educational Research Journal, 19*(2), 60-74.

Jones, C. H., Slate, J. R., Marini, I., & DeWater, B. K. (1993). Academic skills and attitudes toward intelligence. *Journal of College Student Development, 34*, 422-424.

Kirschenbaum, D. S., & Perri, M. G. (1982). Improving academic competence in adults: A review of recent research. *Journal of Counseling Psychology, 25*, 76-94.

Lefcourt, H. M. (1976). *Locus of control: Current trends in theory and research.* New York: Wiley.

Michael, J. L. (1991). A behavioral perspective on college teaching. *The Behavior Analyst, 14*, 229-239.

Munro, B. H (1981). Dropouts from higher education: Path analysis of a national sample. *American Educational Research Journal, 18*, 133-141.

Phares, E. J. (1976). *Locus of control in personality.* Morristown, NJ: General Learning Press.

Schlossberg, N. K., Lynch, A. Q., Chickering, A. W. (1989). *Improving higher education environments for adults: Responsive programs and services from entry to departure.* San Francisco: Jossey-Bass.

Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are effected by quantity of study. *Social Forces, 63*, 945-966.

Schwartz, M. D. (1992). Study sessions and higher grades: Questioning the causal link. *College Student Journal, 26*, 292-299.

Tabachnick, B. G., & Fidell, L. S. (1983). *Using multivariate statistics.* New York: Harper and Row.

Trice, A. (1985). An academic locus of control scale for college students. *Perceptual and Motor Skills, 61*, 1043- 1046.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 1, 63-73

The Harrington-O'Shea Career Decision-Making System (CDM) and the Kaufman Adolescent and Adult Intelligence Test (KAIT): Relationship of Interest Scale Scores to Fluid and Crystallized IQs at Ages 12 to 22 Years

James E. McLean and Alan S. Kaufman
The University of Alabama

Related the six Holland-based Interest Scale scores yielded by the Harrington-O'Shea Career Decision-Making System (CDM) to sex, race, and performance on the KAIT. The sample comprised 254 males and females aged 12 to 22 years. MANOVAs and MANCOVAs (covarying parents' education) were conducted, followed by univariate ANOVAs and ANCOVAs. Sex and Race were significant main effects in the MANOVA, but only Sex was significant in the MANCOVA. The KAIT variables and all interactions were nonsignificant in both multivariate analyses. Follow-up univariate ANCOVAs indicated that males outscored females on the Crafts (Realistic) scale and females outscored males on the Social scale. Race differences in which blacks scored higher than whites and Hispanics on the Business (Enterprising) and Clerical (Conventional) scales failed to achieve significance with parents' education covaried. The present findings were consistent with previous research with the Strong.

The Harrington-O'Shea Career Decision-Making System (CDM; Harrington & O'Shea, 1982) and its recent revision (CDM-R; Harrington & O'Shea, 1992) are interest inventories for assessing career interests and for providing steps that help in understanding oneself and that aid in making effective career decisions. The CDM is intended for students in grades 7-12 and college, and for adults who face decisions involving career change. Like the popular Strong Interest Inventory (Hansen & Campbell, 1985), the CDM provides six scores derived from Holland's hexagonal model of occupational personality types--types that emerge from an interaction among personal and cultural factors with environmental factors (Holland, 1973, 1985). Unlike the Strong and its predecessors (the Strong-Campbell Interest Inventory and Strong Vocational Interest Blank), which have been the subject of scores of research investigations, the CDM and CDM-R have been used to measure vocational interests in relatively few empirical studies (Brown, Ware, & Brown, 1985; Harrington, 1991, 1992; Harrington & O'Shea, 1992, Chapter 8). Nonetheless, because the Strong General Occupational Themes and the CDM scales are both derived directly from Holland's theory, the results of research studies on the CDM (and CDM-R) are able to be interpreted in the context of previous findings on the Strong.

Assessing vocational interests is crucial for assisting individuals with curricular and career choices; such assistance has been a traditional role for counselors and clinical psychologists (Lowman, 1991), and is an emerging role for school psychologists both in academic settings (Bernard & Naylor, 1982; Shepard & Hohenshil, 1983) and in business and industry (Levinson & Shepard, 1986). Empirical research is essential for interest inventories to facilitate their use. The present study was aimed at increasing the empirical foundation of the CDM, an instrument that has been generally well reviewed (Droege, 1984; Manuele-Adkins, 1989; Spitzer & Levinson, 1988) but has been understudied.

This study uses the 1982 CDM rather than the 1992 CDM-R, because the revised version was not available when the present data were gathered. However, the data on the CDM will generalize rather well to Level 2 of the CDM-R, which "closely resembles the original CDM" (Harrington & O'Shea, 1992, p. 1). The data are not generalizable to the new Level 1 of the CDM-R, a less complex version intended for younger students and poor readers.

Specifically, the six Holland-based Interest Scale scores yielded by the CDM were related to the variables of age, sex, race, IQ level, and fluid versus crystallized intelligence for a heterogeneous sample of adolescents and adults. The new Kaufman Adolescent and Adult

---

James E. McLean is a University Research Professor and the Assistant Dean for Research and Service in the College of Education at The University of Alabama. Alan S. Kaufman is a Research Professor of Behavioral Studies in the College of Education at The University of Alabama. Please address correspondence regarding the paper to James E. McLean, Office of Research and Service, The University of Alabama, Tuscaloosa, AL 35487-0231 (Internet: JMCLEAN@BAMAED.UA.EDU).

Intelligence Test (KAIT; Kaufman & Kaufman, 1993), derived from the Horn-Cattell theory of intelligence (Horn, 1985, 1989; Horn & Cattell, 1966, 1967), served as the measure of intelligence level and of the discrepancy between an individual's ability to demonstrate fluid intelligence (solving novel, non-school-related problems) and crystallized intelligence (solving problems that depend on education and acculturation). The relationship between the Horn-Cattell intellectual constructs and the Holland personality constructs is of special interest because several of the Holland-inspired scales on the CDM bear apparent relationships to intellectual functioning. For example, Scientific individuals value mathematics and scientific work very highly and persons who score high on the Clerical scale (called Office Operations on the CDM-R) prefer jobs with clearly defined duties (Harrington & O'Shea, 1982, 1992). Similarly, the writing activities preferred by Arts individuals seem to relate closely to crystallized intelligence, whereas the mechanical preferences of Crafts persons seem more fluid-oriented.

Three recent studies related the same variables investigated in the present study to Holland's themes, as measured by the Strong Interest Inventory, and to the Strong's Basic Interest Scales, for individuals aged 16 to 65 years (Kaufman, Ford-Richards, & McLean, 1993; Kaufman & McLean, 1993; McLean & Kaufman, 1993). The results of the Holland analyses in these studies (using $p < .01$) indicated that: (a) Sex, Race, IQ level, and Fluid-Crystallized discrepancy yielded significant main effects in the multivariate analyses, but Age did not; (b) males scored significantly higher than females on the Realistic and Investigative themes, and females scored significantly higher on the Artistic and Social themes; (c) whites scored significantly higher than blacks on the Realistic theme; (d) the interest profiles of whites and Hispanics were similar to each other, and both differed from profiles for blacks; (e) IQ level related significantly to the Investigative and Artistic themes, with higher IQS associated with higher scores on these themes; (f) Fluid-Crystallized discrepancy did not relate significantly to any of the themes at the .01 level, despite an overall significant main effect in the multivariate analysis; and (g) results of the various analyses were quite consistent, whether or not Educational attainment was covaried.

In this study, the degree to which the aforementioned results with the Strong at ages 16 to 65 years generalized to the CDM at ages 12 to 22 years was explored. In addition, the results of the sex differences on the CDM were compared to the CDM sex differences reported in the manual for grades 7-9, grades 10-12, and college freshmen (Harrington & O'Shea, 1982, Table 6.1). For all samples, the largest sex differences were on the Crafts and Social scales: Males scored about one standard deviation higher than females on the Crafts scale (consistent with the significant difference favoring males on the Strong Realistic theme), and females scored about one standard deviation higher than males on the Social scale (also consistent with the Strong finding).

The relationship between vocational interests and performance on an individually administered, clinical intelligence test is especially important since clinicians commonly administer both types of tests to clients (most typically the Strong and the Wechsler Adult Intelligence Scale--Revised or WAIS-R, Wechsler, 1981; see Harrison, Kaufman, Hickman, & Kaufman, 1988), and interpret such measures jointly in clinical case reports (Lindemann & Matarazzo, 1990; Lowman, 1991).

Method

Subjects

The sample was composed of 254 individuals aged 12 to 22 years (mean age = 15.1 years, SD = 2.0). The group included 134 females (52.8%) and 120 males (47.2%), and was composed of 199 whites (78.3%), 35 blacks (13.8%), and 20 Hispanics (7.9%). Data were also available for 54 additional individuals, but they were eliminated because: (a) the subjects were over-age for the present sample (ages 30 to 58 years; $N = 14$); (b) they were from "other" racial groups (e. g., Asian-Americans, American Indians; $N = 11$); or (c) they had missing data on pertinent variables for the analyses conducted in this study (i.e., age, sex, race; $N = 29$). The older subjects were eliminated to maintain the relative homogeneity of the sample on the variable of age. The "others" were omitted because of their small sample size.

Parents' educational attainment was used to estimate socioeconomic status. Mean number of years of schooling for the total sample was 13.5 years (SD = 2.8), or nearly two years of college or vocational training beyond high school. Mean chronological age and mean years of parental education were as follows for the three racial/ethnic groups: 15.1 years (SD = 2.0) for age and 13.9 years of parents' education (SD = 2.4) for whites; 14.9 years (SD = 1.8) and 13.3 years (SD = 2.6) for blacks; and 15.0 years (SD = 2.4) and 9.4 years (SD = 3.7) for Hispanics. Subjects were tested throughout the United States during the nationwide standardization of the KAIT (Kaufman & Kaufman, 1993).

Instruments

Harrington-O'Shea Career Decision-Making System (CDM). The CDM (Harrington & O'Shea, 1982) was administered for this study. The Strong, intended for students in grades 7-12 and college and for adults contemplating career change, comprises items in a variety of areas. Individuals completing the inventory: (a) select their top two occupational preferences from a list of 18; (b) select their top two school subjects from a list of 14; (c) select one future educational or training plan from a list of nine; (d) select four job values (e. g., security) from a list of 14; (e) select their four strongest abilities from a list of 14; and (f) complete a 120-item interest inventory by responding 0-1-2 to indicate how they feel about each job and job-related activity.

Responses to the latter set of 120 items yield scores on the six Holland-based Interest Scale scores. Each scale is composed of 20 items, and the person's Interest Scale score equals the sum of his or her responses to those 20 items.

The six CDM scales are described as follows (the corresponding Holland personality type appears in parentheses):

Crafts (Realistic)--Generally practical, physically strong, and reserved people who prefer working with tools and objects rather than with words and people, and are interested in concrete, mechanical activities that often involve building things. (Illustrative Crafts occupations: Farmer, Auto mechanic, Dental lab technician)

Scientific (Investigative)--Generally curious, creative, theoretical, studious people who value mathematics and scientific work and often prefer working by themselves. (Illustrative Scientific occupations: Chemist, Architect, Physician)

The Arts (Artistic)--Generally nonconforming, independent, sensitive, self-expressive people who are interested in creative activities such as music, writing, entertainment, and art. (Illustrative Arts occupations: Writer, Interior decorator, Actress)

Social (Social)--Generally popular, sociable, responsible people with strong verbal skills who are concerned with the well-being of others. (Illustrative Social occupations: Social worker, Teacher, Nurse)

Business (Enterprising)--Generally self-confident, energetic, enthusiastic, aggressive people who see themselves as verbally persuasive and are attracted to careers that provide opportunities to lead and persuade others. (Illustrative Business occupations: Salespersons, Government administrators, Stockbrokers)

Clerical (Conventional)--Generally orderly, systematic, dependable people who enjoy organized tasks and the verbal and numerical activities of office work; they prefer occupations in which the duties are clearly defined and often place a high value on financial success and status. (Illustrative Clerical occupations: Secretary, Bank teller, Accountant)

The above definitions refer to "pure" types of each Interest scale, although in reality people are blends of several scales or "types." Interpretation of a person's CDM profile involves examining his or her scores on each of the six Interest scales, and identifying the highest ones. The CDM interpretive system "suggests those career clusters which provide occupational environments consistent with the individual's two highest scores" (Harrington & O'Shea, 1982, p. 23).

CDM norms are provided for grades 7-12 based on a sample of 9,650 males and females stratified on socioeconomic status, and for college students based on a sample of 2,925 males and females stratified by type of institution and form of control (public, private). However, these norms are not used to interpret the Interest Scales, which yield raw scores. Harrington and O'Shea (1982) state: "[I]t is not recommended that interpretation of individual clients' interest inventory results be based on norms. . . . The CDM normative study, however, does afford counselors an opportunity to compare their clients with a carefully selected national sample" (p. 40). By using raw scores, and interpreting scores ipsatively, i.e., comparing a person's scores on each scale to his or her scores on the other scales, the CDM avoids dealing with the common issue of whether to provide separate or combined norms for males and females.

The CDM manual (Harrington & O'Shea, 1982, Tables 7.1-7.3) provides internal consistency and stability data for the Interest Scales. Alpha reliability coefficients for the six CDM scales averaged .92 with a range of .90-.94 for grades 7-9 (N = 4,004); 94 with a range of .90-.94 for grades 10-12 (N = 5,646); and 93 with a range of .90-.94 for college freshmen (N = 2,925). Results were very similar for females and males. Test-retest coefficients over a 5-week interval for 114 high school students and 72 university graduate students yielded median coefficients for the six Interest Scales of .85 for high school females; .80 for high school males; .86 for university females; and .91 for university males. The manual also presents evidence of the CDM's construct, concurrent, and predictive validity (Harrington & O'Shea, 1982, chapter 8). The authors present intercorrelational data among the six scales to provide evidence that the scales measure the intended Holland constructs. Additional evidence for construct validity comes from studies relating the CDM to Holland's and Strong's instruments for assessing the six Holland personality types. Concurrent validity data were presented that compared the CDM codes on Holland's six types with

Holland's own measurement of codes for various occupations for 17 groups of people in diverse occupations; substantial agreement was obtained. Similar results came from a study of college students representing 16 majors. Predictive validity data examined the percentage of subjects whose late 1980 job or educational status agreed with CDM scale scores obtained in the mid-1970s, prior to publication of the CDM. The level of agreement was similar to the level displayed by the Strong inventories in previous studies.

Kaufman Adolescent and Adult Intelligence Test (KAIT). The KAIT (Kaufman & Kaufman, 1993) is a new intelligence test for ages 11 to 85+ years that provides Fluid, Crystallized, and Composite IQS, each with a mean of 100 and standard deviation of 15, and follows the theoretical model of Horn and Cattell (1966, 1967; Horn, 1989). Tasks were developed from the models of Piaget's (1972) formal operations and Luria's (1973) planning ability in an attempt to include high-level, decision-making, adult-oriented tasks. Visual-motor coordination and visual-motor speed are deemphasized, although speed of problem solving is required for several tasks. A Core Battery of six subtests (three Crystallized, three Fluid) yields the three IQS; an Expanded Battery of 10 subtests also includes alternate Crystallized and Fluid subtests and two tasks that measure the delayed recall of information learned previously in the examination. For the present study, only the IQS were used as variables.

The KAIT was normed on 2,000 individuals aged 11 to 85+ years, and was stratified on the variables of age, gender, race or ethnic group, geographic region, and socioeconomic status (parental education for ages 11-24 years, self-education for ages 25 and above). Mean split-half reliability coefficients were .95 for Crystallized IQ, .95 for Fluid IQ, and .97 for Composite IQ. Mean test-retest reliability coefficients, based on 153 normal individuals aged 11-85+ retested after a one-month interval, were as follows: Crystallized IQ (.94), Fluid IQ (.87), and Composite IQ (.94). Exploratory and confirmatory factor analysis supported the construct validity of the Crystallized and Fluid Scales and the placement of subtests on each scale. Correlational analyses with the WISC-R at ages 11-16 ($N = 118$) and WAIS-R at ages 16-83 ($N = 343$) indicated that KAIT Composite IQ correlated in the mid-.80s with Wechsler's Full Scale IQ; KAIT Crystallized and Fluid IQS correlated in the .70s and .80s with Wechsler's Full Scale IQ for these predominantly normal samples.

Procedure

Data for this study were obtained during the nationwide standardization of the KAIT between 1988 and 1991. Qualified examiners who were well trained in the administration and interpretation of individual intelligence tests administered the KAIT. The CDM was self-administered by most standardization subjects aged 12-19 years, although a number of individuals above age 19 were tested as well. For this study, as noted previously, the age range was limited to individuals aged 12 to 22 years. All record forms were machine scored by Consulting Psychologists Press, distributor of the Strong.

Data Analysis

A multiple analysis of variance (MANOVA) was conducted using Race (black-white-Hispanic), Sex, and IQ level as independent variables, and scores on the six CDM Interest Scales as dependent variables. The total sample was divided into three levels of intelligence: 110-160 ($N = 73$); 90-109 ($N = 125$); 40-89 ($N = 56$). The MANOVA was followed by six univariate ANOVAs, one for each Interest Scale.

Next, a MANCOVA was conducted using Race, Sex, and Fluid-Crystallized IQ discrepancy on the KAIT as independent variables and scores on the six CDM Interest Scales as dependent variables; parents' educational attainment was the covariate. The total sample was divided into three Fluid (F)-Crystallized (C) discrepancy categories: $F > C$ ($N = 59$); $F = C$ ($N = 142$); and $C > F$ ($N = 53$). The average Fluid-Crystallized IQ discrepancy required for statistical significance at the .05 level for the total KAIT standardization sample is 9 points (Kaufman & Kaufman, 1993), so differences of at least 9 points in favor of Fluid IQ were needed to classify a person as $F > C$; differences of at least 9 points in favor of Crystallized IQ were needed to classify a person as $C > F$; and differences of + 8 points classified a person as $F = C$. The MANCOVA was followed by six univariate ANCOVAs, one for each Interest Scale.

Significant *F* values in the ANOVAs and ANCOVAs were followed up with Tukey's Honestly Significant Differences (HSD) test to determine the significance of differences between pairs of means.

Educational attainment was used as a covariate in the second set of analyses, but it was undesirable to use it in the first set because education and intelligence are so closely correlated (Kaufman, 1990, Chapter 6); any control for education in the initial analyses would have compromised interpretation of the relationship of intelligence level to interests.

An alpha level of .05 was used for all multivariate analyses, but more stringent criteria were used in the univariate analyses ($p < .01$) to offset the chance factors that accompany conducting multiple analyses simultaneously.

Results and Discussion

The results of the MANOVA and MANCOVA are summarized in Table 1 and Table 2, respectively. Sex ($p < .001$) and Race ($p < .05$) were significant main effects in the MANOVA, but only Sex ($p < .001$) remained significant with Parents' education covaried. IQ level and Fluid-Crystallized discrepancy failed to reach significance in either multivariate analysis, and all interactions were nonsignificant.

Table 1
Wilks Lambda and F Statistics for Each Main Effect and Interaction in the MANOVA of the Six Interest Scales of the Harrington-O'Shea Career Decision-Making System

Variable	Wilks Lambda	*F*
Sex	.801	9.63***
Race (Black/White/Hispanic)	.910	1.87*
IQ Level (40-89, 90-109, 110-160)	.926	1.51
Sex × Race	.966	0.67
Sex × IQ	.946	1.09
Race × IQ	.906	0.97
Sex × Race × IQ	.957	0.58

$^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table 2
Wilks Lambda and F Statistics for Each Main Effect and Interaction in the MANCOVA of the Six Interest Scales of the Harrington-O-Shea Career Decision-Making System

Variable	Wilks Lambda	*F*
Sex	.783	10.69***
Race (Black/White/Hispanic)	.932	1.38
Fluid (F)-Crystallized (C) Discrepancy	.941	1.19
Sex × Race	.964	0.71
Sex × F-C	.950	0.99
Race × F-C	.926	0.75
Sex × Race × F-C	.954	0.61

$^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

Note. Fluid (F)-Crystallized (C) Discrepancy equals:
F > C: Fluid IQ significantly (9+ pts.) higher than Crystallized IQ ($p < .05$)
F = C: Fluid not significantly different from Crystallized IQ--less than 9 points difference in either direction
C > F: Crystallized IQ significantly (9+ pts.) higher than Fluid IQ ($p < .05$)

Table 3 shows the *F* values for the Sex and Race main effects in the univariate ANOVAs and ANCOVAs; as indicated, both variables were significant in the MANOVA, and Sex was also significant in the MANCOVA. The *F* values for Race from the ANCOVAs should not be interpreted, but are presented in Table 3 to provide an indication of the effect of covarying parents' education on each Interest Scale. Table 3 also shows the *F* value for the covariate in each ANCOVA. The covariate of parents' education was significantly ($p < .01$) related to three Interest Scales (Scientific, The Arts, Business), supporting the advisability of covarying socioeconomic status in this study.

Although the .01 level is used to denote significant findings for the univariate analyses, significance at $p < .05$ is nonetheless indicated in Table 3 both for informational purposes and to permit readers to apply alternate alpha levels.

Table 4 presents means and SDs for the total sample and for each subsample to aid in the interpretation of the ANOVAs and ANCOVAs.

Sex Differences

Crafts, associated with Holland's Realistic theme, was a significant main effect ($p < .001$) in both the ANOVA and ANCOVA, with males scoring about 1 SD higher than females. Females scored significantly higher than males at the .01 level on the Clerical Scale in the ANOVA and on the Social Scale in the ANCOVA. Since parental education, an estimate of socioeconomic status, contributes unwanted variance in the male-female comparisons, the significant difference on Social is interpreted as reflecting a true sex-related difference, but the Clerical difference is interpreted as primarily a function of socioeconomic differences between males and females (mean parents' education of 13.8 years and 13.1 years, respectively).

The significant sex differences on Crafts, favoring males, and on Social, favoring females, conforms precisely to the most prominent sex differences reported in the CDM manual (Harrington & O'Shea, 1982, Table 6.1), and also in the CDM-R manual (Harrington & O'Shea, 1992, Table 6.1). These findings are likewise consistent with recent findings on the Strong Interest Inventory that showed males outscoring females on Holland's Realistic theme and females outscoring males on the Social theme (McLean & Kaufman, 1993). In that study, males also scored higher on the Investigative theme and females scored higher on the Artistic theme. In the present study, males scored 6 points (about ½ SD) higher than females on the Scientific Scale, which is intended to reflect Holland's Investigative theme. Although that difference reached significance at the .05 level, the .01 level was needed for this study; further, the difference failed even to reach significance at the .05 level when socioeconomic status was covaried in the ANCOVA, and the CDM and CDM-R manuals reveal apparently trivial mean male-female differences on the Scientific Scale (Harrington & O'Shea, 1982, 1992).

Table 3
Univariate *F* Values for Variables that Were Statistically Significant in the MANOVA or MANCOVA, for the Education Covariate

Variable	Crafts	Scientific	The Arts	Social	Business	Clerical
ANOVA						
Sex	21.31***	4.92*	0.10	6.30*	1.89	8.70**
Race	0.22	0.72	2.57	3.65*	5.86**	7.13***
ANCOVA						
Sex	29.93***	3.33	0.16	7.25**	0.68	2.94
Race	1.46	0.27	0.32	3.33*	3.01	4.17*
Parents' Education	4.34*	7.48**	8.05**	2.01	7.61**	0.31

Note: Main effects are described in the Table 1 footnotes. The Race main effect was significant in the MANOVA but not the MANCOVA. Univariate *F* values are presented here for the ANCOVAs merely for informational purposes; they should not be interpreted.

Table 4
Means and Standard Deviations of Raw Scores on the Six Interest Scales of the Harrington-O'Shea Career Decision-Making System, by Sex, Race, KAIT Composite IQ, and KAIT Fluid-Crystallized Discrepancy (*N* = 254)

	Crafts		Scientific		The Arts		Social		Business		Clerical	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Sex												
Female	6.0	7.8	11.6	11.4	17.5	10.8	20.7	9.9	16.9	9.6	15.1	10.4
Male	14.5	9.9	17.6	11.8	15.2	9.8	13.4	9.6	14.7	9.4	11.0	9.2
Race/Ethnic												
White	9.7	9.8	14.7	12.3	16.7	10.7	16.4	10.5	15.3	9.6	12.5	9.9
Black	12.3	10.8	14.9	11.2	18.4	8.0	22.9	8.4	21.8	8.5	19.3	9.7
Hispanic	9.4	8.2	10.8	8.8	10.4	8.7	15.4	9.9	11.2	7.0	8.8	8.0
IQ Level												
110-160	10.0	9.8	17.9	12.7	18.5	11.0	16.9	9.4	17.2	9.0	13.9	10.1
90-109	9.1	9.2	13.2	11.8	16.1	10.2	16.6	11.0	15.6	10.0	12.8	10.3
40-89	12.1	10.9	12.5	10.0	14.5	9.6	19.0	10.1	14.7	9.3	13.1	9.6
Fluid/ Cryst.												
C > F	10.6	10.7	14.8	11.8	18.6	9.3	18.2	10.5	16.2	9.7	12.3	9.8
F = C	9.8	9.3	14.8	12.0	16.9	10.8	17.2	10.2	17.4	9.5	14.2	10.3
F > C	10.0	10.4	13.1	12.0	13.3	9.6	16.5	10.7	11.9	8.7	11.3	9.4
Total	10.0	9.8	14.4	11.9	16.4	10.4	17.2	10.4	15.9	9.6	13.1	10.0

Note: Fluid/Cryst. = Fluid-Crystallized IQ Discrepancy. Sample sizes are as follows: Sex: Female (*N* = 134); Male (*N* = 120). Race: Whites (*N* = 199); Blacks (*N* = 35); Hispanics (*N* = 20). IQ: 110-160 (*N* = 73); 90-109 (*N* = 125); 40-89; (*N* = 56). F-C: C > F (*N* = 53); F = C (*N* = 142); F > C (*N* = 59).

Strong research, in general, has consistently identified higher scores by males on scales akin to Realistic/Crafts and Investigative/Scientific (Apostal & Marks, 1990; Ford-Richards, 1992; Hansen & Campbell, 1985; Lapan, Boggs, & Morrill, 1989), but the literature reveals inconsistent results regarding sex differences favoring females on Social scales. Some research suggests no meaningful difference on interests pertaining to the Social theme (Apostal & Marks, 1990; Hansen & Campbell, 1985); other studies suggest that females have higher Social codes than men or are more likely to choose social-oriented occupations (Hecht, 1980; Smart, 1989). When significant sex differences have occurred on the Strong social theme, these differences have tended not to be unusually large in magnitude (e. g., McLean & Kaufman, 1993).

The large difference favoring females on the Social scale in this study, in the CDM normative sample (Harrington & O'Shea, 1982), and in the CDM-R normative sample (Harrington & O'Shea, 1992), suggests that the CDM/CDM-R Social Scale is more likely than the Strong Social theme to produce meaningful differences in favor of females. Analogously, the Strong Investigative theme seems more likely than the CDM/CDM-R Scientific Scale to produce higher scores by males than females.

Race Differences

Race differences achieved significance at the .01 level for the Business (Enterprising) and Clerical (Conventional) Scales, and fell just short at $p < .05$ on the Social Scale. Follow-up Tukey HSD analyses indicated that blacks scored significantly higher than both whites and Hispanics on all three of these scales, but the means for whites and Hispanics did not differ significantly from each other (see Table 4 for means and SDs by subsample). In the MANCOVA, however, the main effect for Race failed to reach significance, precluding interpretation of the univariate analyses for this variable. Examination of these F values, presented in Table 3 for comparative purposes, indicates that none of the Interest Scales was significant at the .01 level when socioeconomic status was covaried. The notable difference between the findings for the MANOVA/ANOVAs versus the MANCOVA/ANCOVAs suggests that the apparent racial differences on the CDM were primarily a function of the different socioeconomic backgrounds for the three groups: Mean number of years of schooling for parents were about 14, 13, and 9, respectively, for whites, blacks, and Hispanics.

Some similarities are noted between the present results and the results of the studies of racial differences on the Strong (Kaufman et al., 1993; Kaufman & McLean, 1993). Blacks scored higher than whites on the Social, Enterprising, and Conventional themes in the ANOVAs (Kaufman et al., 1993), the same three Holland-related CDM Interest Scales that produced the largest differences in favor of blacks in this study's ANOVAs. Another consistency is the finding that the Hispanics scored fairly similarly to whites, and both groups differed from blacks, in their interest profiles (Kaufman & McLean, 1993). This finding is noteworthy because it occurred despite the fact that the whites were far more similar to the blacks in socioeconomic status, both in this study and in the Strong investigation, than they were to Hispanics.

The main differences in the two studies are: (a) Whites and Hispanics scored higher than blacks on the Realistic and Investigative themes on the Strong, with a striking difference on Realistic, but analogous differences on the CDM Crafts and Scientific Scales were not found; and (b) Race was a significant variable in the Strong multivariate analyses both with and without a socioeconomic covariate, but the significant Race difference on the CDM disappeared when parents' education was covaried. With the Strong, the significant difference favoring blacks on the Conventional Scale was no longer significant in the ANCOVA, although the other significant differences remained, in the study of black-white differences (Kaufman et al., 1993). When Hispanics were included in the analysis (Kaufman & McLean, 1993), significant Race differences were obtained in the ANOVAs for the Realistic, Investigative, and Artistic themes; with socioeconomic status covaried, these three themes were significant, and they were joined by a fourth, the Social theme. Other investigations of various Strong inventories have likewise supported substantial differences in the interest profiles for blacks versus whites (Carter & Swanson, 1990; Hines 1983/1984; Swanson, 1992; Whetstone & Hayles, 1975; Yura,1985/1986), and these differences have been congruent with the findings of the Kaufman et al. (1993) and Kaufman and McLean (1993) studies.

The prevalence and consistency of the racial differences in previous studies of the Strong impelled Kaufman et al. (1993) to propose separate norms on the Strong for blacks and whites to foster a fairer and less stereotypical interpretation of the interest profiles of black men and women. The present study with the CDM does not agree with the previous Strong findings for blacks versus whites, so no proposal of separate CDM or CDM-R norms for blacks and whites is warranted. However, the relatively small samples of blacks ($N = 35$), and especially of Hispanics ($N = 20$), makes the present results tentative pending future investigation.

Similarly, the results of the previous Strong/KAIT studies indicated that Race was a much more important variable than Educational attainment in determining an individual's pattern of interests on the Strong (Kaufman et al., 1993; Kaufman & McLean, 1993); covarying socioeconomic status had some impact on the pattern of racial differences, but the bulk of significant differences remained even with the covariate. In the present CDM study, the reverse was true: Parents' education was apparently a more important variable than Race in determining a person's interest profile, a finding that supports Slaney's (1980; Slaney & Brown, 1983) contention that socioeconomic variables must be fully taken into account when investigating race differences in vocational interests. Again, however, the relatively small samples of minority individuals in the present study mitigate the present results, and make them hypotheses for follow-up investigation.

Differences on the KAIT

In the recent KAIT/Strong investigation, the variable of KAIT IQ level was a significant main effect in the MANOVA and in the univariate ANOVAs for the Investigative and Artistic themes (McLean & Kaufman, 1993); intelligence level was directly related to individuals' scores on these two themes. The results for Investigative are consistent with: (a) the Investigative person's interest in science and in solving abstract problems; (b) with a variety of findings with various Strong inventories (Lowman, 1991); (c) with the fact that its development is related to educational attainment, undergraduate grades, and socioeconomic status (Smart, 1989); and (d) mental ability is the most predictive variable of success for Investigative occupations (Brody, 1985). The relationship between IQ and the Artistic theme has not typically been found in research investigations (Lowman, 1991).

The Scientific Scale of the CDM/CDM-R, derived from Holland's Investigative theme, bears the same logical relationship to intelligence; "Scientific persons value mathematics and scientific work [and] . . . tend to be curious, creative, theoretical, and studious (Harrington & O'Shea, 1992, p. 4). As indicated in Table 4, the mean for people with IQS > 110 was about ½ SD higher than the mean for those with IQS < 89, but that trend was not significant in this study. The Scientific Scale did, however, relate significantly ($p < .001$) to the socioeconomic covariate of Parents' education (see Table 3).

The present findings suggest that although the variable of socioeconomic status should be taken into account when interpreting a person's CDM profile, the variable of intelligence need not be considered. The Fluid-Crystallized discrepancy likewise does not affect a person's interest profile on the CDM, and provides no additional information for assisting counselors and psychologists in the interpretation of a client's Interest Scale profile. With the Strong, clinicians were advised by McLean and Kaufman (1993) to take IQ level into account when interpreting Holland's themes; but, similar to CDM results, the Fluid-Crystallized discrepancy did not relate meaningfully to any of Holland's six themes as measured by the Strong Interest Inventory (McLean & Kaufman, 1993), and does not facilitate the counseling process.

Conclusions

Scores on the CDM Interest Scales did not relate to most of the variables studied, especially when socioeconomic status was covaried. The sex differences observed are consistent with previous research on the CDM, CDM-R, and Strong. The lack of racial differences when parents' education was covaried, though possibly a function of the small subsamples of blacks and Hispanics, differs from the bulk of Strong research; if replicated, this finding may imply a real difference between the CDM and Strong regarding black-white profile differences on vocational interests. That possibility has important implications for the counseling process because the history of racial discrimination in the United States has accentuated the importance of race in the labor market and in the occupational structure of this country (Smith, 1983). The differential treatment of blacks in the labor market has been both class-bound and race-bound. There has been some occupational mobility for blacks, but Smith reported that black women have moved slower than black men out of their previous low status occupations. The range of occupational choices may be limited for blacks, either by actual or perceived unavailability of various types of occupations (Dawkins, 1989). These diverse factors conceivably contribute to the different interest patterns observed on the Strong and related instruments. If differential patterns are found not to exist on the CDM/CDM-R for blacks versus whites--but instead exist for people from different socioeconomic backgrounds--then that finding would suggest that racial membership need not influence the counselor's vocational guidance suggestions when the Career-Decision-Making System is used.

Despite a striking difference in educational attainment for whites and Hispanics in this study, their profiles of interests were fairly congruent. In contrast, the interest profiles of both Hispanics and whites differed notably from the profile for blacks on the CDM Interest Scales These findings parallel the results of the recent Strong study (Kaufman & McLean, 1993), and of another recent investigation (Kaufman, Kaufman, & McLean, 1993) of the Myers-Briggs Typology Inventory (Myers & McCaulley, 1985). Also, the lack of significant Sex × Race interactions in this CDM study, and in the previous Strong and Myers-Briggs investigations, suggests that the findings for Hispanics, blacks, and whites generalize to both males and females.

The KAIT intellectual variables did not relate significantly to CDM Interest Scale raw scores in this study; the lack of significance for intellectual level, in particular, differs from previous results with the Strong (McLean & Kaufman, 1993). If this finding is replicated, then CDM/CDM-R interpretation may not be affected by intelligence level or pattern of ability, an important bit of information because clinicians commonly administer both intelligence tests and interest inventories to clients seeking vocational guidance (Harrison et al., 1988; Lindemann & Matarazzo, 1990; Lowman, 1991).

References

Apostal, R., & Marks, C. (1990). Correlations between the Strong-Campbell and Myers-Briggs Scales of Introversion-Extraversion and career interests. *Psychological Reports, 66*, 811-816.

Bernard, M. E., & Naylor, F. D. (1982). Vocational guidance consultation in school settings. In T. R. Kratochwill (Ed.), *Advances in school psychology* (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum.

Brody, N. (1985). The validity of tests of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 353-389). New York: Wiley.

Brown, D., Ware, W. B., & Brown, S. T. (1985). A predictive validation of the Career Decision-Making System. *Measurement and Evaluation in Counseling and Development, 18*, 81-85.

Carter, R. T., & Swanson, J. L. (1990). The validity of the Strong Interest Inventory with black Americans: A review of the literature. *Journal of Vocational Behavior, 36*, 195-209.

Dawkins, M. P. (1989). The persistence of plans for professional careers among blacks in early adulthood. *Journal of Negro Education, 58*, 220-231.

Droege, R. C. (1984). The Harrington-O'Shea Career Decision-Making System. In D. Keyser & R. Sweetland (Eds.), *Test critiques* (Vol. 1, p. 326). Kansas City, MO: Test Corporation of America.

Ford-Richards, J. M. (1992). *A comparison of the general occupational theme scores of black Americans and white Americans on the Strong Interest Inventory.* Unpublished doctoral dissertation, The University of Alabama.

Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII* (4th ed.). Stanford, CA: Stanford University Press (Distributed by Consulting Psychologists Press).

Harrington, T. F. (1991). The cross-cultural applicability of the career decision-making system. *The Career Development Quarterly, 39*, 209-220.

Harrington, T. F. (1992). The concurrent validity of the career decision-making system cross-culturally. *International Journal for the Advancement of Counseling, 15*, 39-45.

Harrington, T. F., & O'Shea, A. J. (1982). *Manual for the Harrington-O'Shea Career Decision-Making System.* Circle Pines, MN: American Guidance Service.

Harrington, T. F., & O'Shea, A. J. (1992). *Manual for the Harrington-O'Shea Career Decision-Making System-Revised.* Circle Pines, MN: American Guidance Service.

Harrison, P. L., Kaufman, A. S., Hickman, J. A., & Kaufman, N. L. (1988). A survey of tests used for adult assessment. *Journal of Psychoeducational Assessment, 6*, 188-198.

Hecht, A. B. (1980). Nursing career choice and Holland's theory: Are men and blacks different? *Journal of Vocational Behavior, 16*, 208-211.

Hines, H. (1984). The Strong-Campbell Interest Inventory: A study of its validity with a sample of black college students (Doctoral dissertation, University of Maryland, 1983). *Dissertation Abstracts International, 45*, 1901B.

Holland, J. L. (1973). *Making vocational choices: A theory of careers.* Englewood Cliffs, NJ: Prentice-Hall.

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments.* Englewood Cliffs, NJ: Prentice-Hall.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 267-300). New York: Wiley.

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research.* New York: W. H. Freeman.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253-270.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107-129.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston, MA: Allyn and Bacon.

Kaufman, A. S., Ford-Richards, J. M., & McLean, J. E. (1993). *Black-white differences on the Strong Interest Inventory general occupational themes and basic interest scales at ages 16 to 65.* Manuscript submitted for publication.

Kaufman, A. S., & Kaufman, N. L. (1993). *Manual for the Kaufman Adolescent and Adult Intelligence Test (KAIT).* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Kaufman, N. L., & McLean, J. E. (1993). Profiles of Hispanic adolescents and adults on the Myers-Briggs Typology Inventory. *Perceptual and Motor Skills, 76,* 628-630.

Kaufman, A. S., & McLean, J. E. (1993). *Profiles of Hispanic adolescents and adults on the Holland Themes and Basic Interest Scales of the Strong Interest Inventory.* Manuscript submitted for publication.

Lapan, R. T., Boggs, K. R., & Morrill, W. H. (1989). Self-efficacy as a mediator of investigative and realistic general occupational themes on the Strong-Campbell Interest Inventory. *Journal of Counseling Psychology, 36,* 176-182.

Levinson, E. M., & Shepard, J. W. (1986). School psychology in business and industry. *Psychology in the Schools, 23,* 152-157.

Lindemann, J. E., & Matarazzo, J. D. (1990). Assessment of adult intelligence. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed.) (pp. 70-101). New York: Pergamon.

Lowman, R. L. (1991). *The clinical practice of career assessment.* Washington, DC: American Psychological Association.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology.* New York: Basic Books.

Manuele-Adkins, C. (1989). Review of the Harrington-O'Shea Career Decision-Making System. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurements yearbook* (pp. 344-345). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.

McLean, J. E., & Kaufman, A. S. (1993). *The Strong Interest Inventory and the Kaufman Adolescent and Adult Intelligence Test (KAIT): Relationship of general occupational themes and basic interest scales to IQ level and fluid-crystallized discrepancy at ages 16 to 65 years.* Manuscript submitted for publication.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15,* 1-12.

Shepard, J., & Hohenshil, T. H. (1983). Career development functions of practicing school psychologists: A national study. *Psychology in the Schools, 20,* 445-449.

Slaney, R. B. (1980). An investigation of racial differences on vocational variables among college women. *Journal of Vocational Behavior, 16,* 197-207.

Slaney, R. B., & Brown, M. T. (1983). Effects of race and socioeconomic status on career choice variables among college men. *Journal of Vocational Behavior, 23,* 257-269.

Smart, J. C. (1989). Life history influences on Holland vocational type development. *Journal of Vocational Behavior, 34,* 69-87.

Smith, E. J. (1983). Issues in racial minorities' career behavior. In W. B. Walsh & S. H. Osipow (Eds.), *Handbook of vocational psychology* (Vol.1). Hillsdale, NJ: Lawrence Erlbaum.

Spitzer, D., & Levinson, E. M. (1988). A review of selected vocational interest inventories for use by school psychologists. *School Psychology Review, 17,* 673-692.

Swanson, J. L. (1992). The structure of vocational interests for African-American college students. *Journal of Vocational Behavior, 40,* 144-157.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale--Revised.* San Antonio: The Psychological Corporation.

Whetstone, R. D., & Hayles, V. R. (1975). The SVIB and black college men. *Measurement and Evaluation in Guidance, 8,* 105-109.

Yura, C. A. (1986). The Strong-Campbell Interest Inventory: An investigation of black college students and white college students on the Strong-Campbell Interest Inventory. (Doctoral dissertation, West Virginia University, 1985). *Dissertation Abstracts International, 46,* 2572A. (Doctoral dissertation, University of Maryland, 1983). *Dissertation Abstracts International, 45,* 1901B.

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◇◇

Please enter a subscription to Research in the Schools for:

Name: ______________________________

Institution: ______________________________

Address: ______________________________

____

			COST
Individual Subscription ($25 per year)	Number of years	_____	_______
Institutional Subscription ($30 per year)	Number of years	_____	_______
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	_____	_______
TOTAL COST:			_______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
Research in the Schools
The University of Alabama
P. O. Box 870231
Tuscaloosa, AL 35487-0231

Please note that a limited number of copies of Volume 1 are available and can be purchased for the same subscription prices noted above.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

* * *

MSERA Membership/Renewal Form

(Please print or type)

NAME: ____________________

TITLE: ____________________

INSTITUTION: ____________________

MAILING ADDRESS: ____________________
____
____

PHONE: ____________________ FAX ____________________

ELECTRONIC MAIL ADDRESS: ____________________

MSERA MEMBERSHIP: New___ Renewal ___
ARE YOU A MEMBER OF AERA? Yes___ No___
WOULD YOU LIKE INFORMATION ON AERA MEMBERSHIP? Yes___ No___

DUES:	Professional	$15.00	______
	Student	$10.00	______

VOLUNTARY TAX DEDUCTIBLE CONTRIBUTION
TO MSER FOUNDATION ______

TOTAL ______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. Dorothy D. Reed (MSERA)
Headquarters, Air University
USAF, 55 LeMay Plaza South
Maxwell AFB, AL 36112-6335

Mid-South Educational Research Association
and The University of Alabama
Post Office Box 870231
Tuscaloosa, AL 35487-0231

NON-PROFIT ORG.
U.S. POSTAGE
PAID
TUSCALOOSA, AL
PERMIT NO. 16

TMC-1SE

MSERA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and The University of Alabama at Birmingham.

Volume 2, Number 2 **Fall 1995**

School Transformation Through Invitational Education 1
William Watson Purkey and David Strahan

The Effect of Random Class Assignment on Elementary Students' Reading and Mathematics Achievement 7
Jayne B. Zaharias, C. M. Achilles, and Van A. Cain

Retention Across Elementary Schools in a Midwestern School District 15
Sim Gurewitz and Jack Kramer

Using Research Results on Class Size to Improve Pupil Achievement Outcomes 23
C. M. Achilles, Patrick Harman, and Paula Egelson

Biology Students' Beliefs about Evolutionary Theory and Religion 31
Anne Sinclair and Beatrice Baldwin

Principal Leadership Style, Personality Type, and School Climate 39
Dawn T. Hardin

Preservice Teachers and Standardized Test Administration: Their Behavioral Predictions Regarding Cheating. 47
Karyn Wellhousen and Nancy K. Martin

A Typology of School Climate Reflecting Teacher Participation: A Q-technique Study 51
Dianne L. Taylor, Bruce Thompson, and Ira E. Bogotch

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1995 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 2, Number 2 **Fall 1995**

EDITORS
James E. McLean, *The University of Alabama at Birmingham*
Alan S. Kaufman, *Psychological Assessment Resources, Inc. (PAR)*

PRODUCTION EDITOR
Margaret L. Glowacki, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Michèle Carlier, *University of Reims, Champagne Ardenne (France)*
Sheldon B. Clark, *Oak Ridge Institute for Science and Education*
Michael Courtney, *Henry Clay High School (Lexington, KY)*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Kentucky*
Donald F. DeMoulin, *Western Kentucky University*
R. Tony Eichelberger, *University of Pittsburgh*
Daniel Fasko, Jr., *Morehead State University*
Patrick Ferguson, *Arkansas Tech University*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Samuel Hinton, *Eastern Kentucky University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
Randy W. Kamphaus, *University of Georgia*
Jwa K. Kim, Middle *Tennessee State University*
Jimmy D. Lindsey, *Southern University and A & M College*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Peter Melchers, *University of Cologne (Germany)*
Claire Meljac, *Psychologue AU C.H.S. Sainte-Anne (France)*
Soo-Back Moon, *Hyosung Women's University (Korea)*
Arnold J. Moore, *Mississippi State University*
Thomas D. Oakland, *University of Texas*
William W. Purkey, *University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Clemson University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Arkansas State University*
Bruce Thompson, *Texas A & M University*
Anne G. Tishler, *The University of Montevallo*
Wayne J. Urban, *Georgia State University*

GRADUATE STUDENT EDITORIAL BOARD
Vicki Benson, *The University of Alabama*
Ann T. Georgian, *The University of Southern Mississippi*
Jin-Gyu Kim, *The University of Alabama*
Robert T. Marousky, *University of South Alabama*
Jerry G. Mathews, *Mississippi State University*
Dawn Ossont, *Auburn University*
Malenna A. Sumrall, *The University of Alabama*

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 1-6

School Transformation Through Invitational Education

William Watson Purkey and David Strahan
The University of North Carolina at Greensboro

The article describes Invitational Education theory, illustrates it with an analogy, and describes its application with seven middle schools. Invitational Education is based on five P's--people, places, policies, processes, and programs. The Invitational Education Theory is the interrelationships among these five P's and the process by which they are implemented. The five P's are illustrated using a starfish analogy. The Invitational Education G.O.A.L.S. Process is the method by which the five P's are implemented. Examples of the process are illustrated to be markedly different than the "factory" model typically used. It is concluded that the schools should become more like families than like factories.

> Invitational Education has a much wider focus of application than is typically discussed in other self-theories. It is deliberately aimed at broader goals than students and their achievement alone. It is geared to the total development of all who interact within the school. It is concerned with more than grades, attendance and even perceptions of self. It is concerned with the skills of becoming.
>
> William Stafford,
> *The Forum, 11, 3*
> December, 1990

Every school we have visited lately seems to be in the midst of some form of school improvement process. Some schools are "restructuring." Others are "reforming." Still others are complying with state and local mandates to "write up a plan to do better." Some of these efforts are beginning to demonstrate meaningful improvement in the quality of life in schools. Others seem to be more mechanical in nature. The differences between meaningful and mechanical efforts may have much to do with the processes employed. Any process that lacks theoretical underpinnings is unlikely to succeed in school. This article describes a successful school transformation process based on "Invitational Education," a theoretical model derived from invitational theory (Purkey & Novak, 1996).

Invitational theory consists of a body of assumptions offered to understand those myriad signal systems that exist in the human environment and that offer something beneficial for consideration and adoption. It is proposed as a theory of practice for communicating caring and appropriate messages that are intended to summon forth the realization of human potential.

The Invitational Education approach to school transformation offered in this article has been developed in collaboration with teachers and administrators from seven different middle-level schools who have worked with us in three research demonstrated projects sponsored by the Z. Smith Reynolds Foundation, the RJR Nabisco STAR Project, and the International Alliance for Invitational Education.

Overview

Invitational Education (Purkey & Novak, 1988, 1996) provides a framework for thinking about who we are and what we hope to accomplish in education. The basic goal of Invitational Education is to create a total school environment that intentionally summons success for everyone associated with the school. Four basic premises undergird Invitational Education:

1. Education is a cooperative, collaborative activity where process is as important as product.
2. People are able, valuable, and responsible and should be treated accordingly.
3. People possess untapped potential in all areas of human endeavor.
4. Human potential can best be realized by places, policies, processes, and programs specifically designed to invite development, and by people who are intentionally inviting with themselves and others, personally and professionally (Purkey & Novak, 1988, pp. 12-13).

William Watson Purkey is Professor of Counselor Education, School of Education, The University of North Carolina at Greensboro and Co-Director of the International Alliance for Invitational Education. David Strahan is Associate Professor of Curriculum and Instruction at The University of North Carolina at Greensboro. The authors may be contacted at the School of Education, The University of North Carolina at Greensboro, Greensboro, NC 27412.

These basic premises provide a guiding theory for school transformation. Invitational Education begins with the fundamental belief that ALL students can succeed. While almost every school has witnessed increasing attention to meeting the needs of "at-risk" students, Invitational Education insists that meeting the needs of all students is the best way to meet the needs of any one student. The goal is to make schooling a more exciting, satisfying, and enriching experience for everyone - all students, all teachers, all visitors. Everything addressed in developing an agenda for improvement should contribute to a more inviting "zeitgeist" or "spirit" within the school. Such an effort goes far beyond "restructuring" or even "reforming." What is at stake is the transformation of the school.

In "Reframing Reform," Terrance Deal (1990) analyzes the failure of many attempts to "reform" education and makes a compelling case for "transformation." He notes that

> reform focuses on correcting weaknesses in existing practices, a focus that is often reduced to tinkering with structural features: revising the schedule, designing a year-round school year, planning tutoring programs, recruiting more parent volunteers. Transformation, in contrast, addresses alterations in fundamental character.

If we are to transform schools, it is important to acknowledge that schools are "complex social organizations held together by a symbolic webbing" rather than "formal systems driven by goals, official roles, commands, and rules" (p. 7). In this respect, "the core problems of schools are more spiritual than technical" (Deal, 1990, p. 12).

By focusing on human potential and collaborative processes, Invitational Education provides a vehicle for transformation. Oberg (1987), Strahan (1990), and others have found that educators' decisions are shaped by their basic "orientations" toward themselves, schooling, and their notions of "the good" that frame their personal and instructional decisions. How we see ourselves and our students, how we view the nature of "good" teaching and learning, how we think about schools and schooling—these are notions that shape our decisions as educators. In doing so, these orientations help create our classroom climate, and, when shared, shape school culture.

How Invitational Education Works

The "five P's" of Invitational Education, standing for *p*eople, *p*laces, *p*olicies, *p*rocesses, and *p*rograms, provide the means for developing more explicit notions of the "good" (Purkey & Novak, 1996). One of the basic premises of Invitational Education is that human potential can best be realized by *people* who are intentionally inviting with themselves and others, personally and professionally, and by *places, policies, processes*, and *programs* specifically designed to invite development. These five powerful "P's" address the global nature of the school and seek to transform the educative process by applying steady and continuous pressure from a number of points . . . much like the starfish conquers oysters.

The Starfish Analogy

The starfish lives to eat oysters. To defend itself, the oyster has two stout shells that fit tightly together and are held in place by a powerful muscle. When a starfish locates an oyster, it places itself on the top shell. Then gradually, gently and continuously, the starfish uses each of its five points in turn to keep steady pressure on the one oyster muscle. While one point works, the others rest. The single oyster muscle, while incredibly powerful, gets no rest. Inevitably and irresistibly, the oyster shells open, and the starfish has its meal. Steady and continuous pressure from a number of points can overcome the biggest muscles of oysters, and by analogy, the biggest challenges in schools, such as school safety, and the largest obstacles to school transformation, such as apathy and lack of common cause. Here is how the invitational education starfish analogy looks when the "Five-P" approach is applied to school improvement. (See Figure 1.)

The Five Powerful P's

People

In planning efforts that improve the quality of life for the *PEOPLE* of the school, we can ask ourselves how we see ourselves and our students, how we envision our relations with each other, and how we can extend and nurture those caring relationships in ways that summon forth human potential.

Places

In considering improvements regarding *PLACES*, we can examine our facilities and grounds and find ways to enhance the total physical environment of the school. Is this a place where people want to be and want to learn?

Policies

In reviewing our *POLICIES*, we can identify rules and regulations that may be "disinviting" and find ways to make them more inclusive, encouraging, and involving. Given the importance of the language we use to describe our operations and expectations, policies become critical "semantic webs" that shape the spirit of the school.

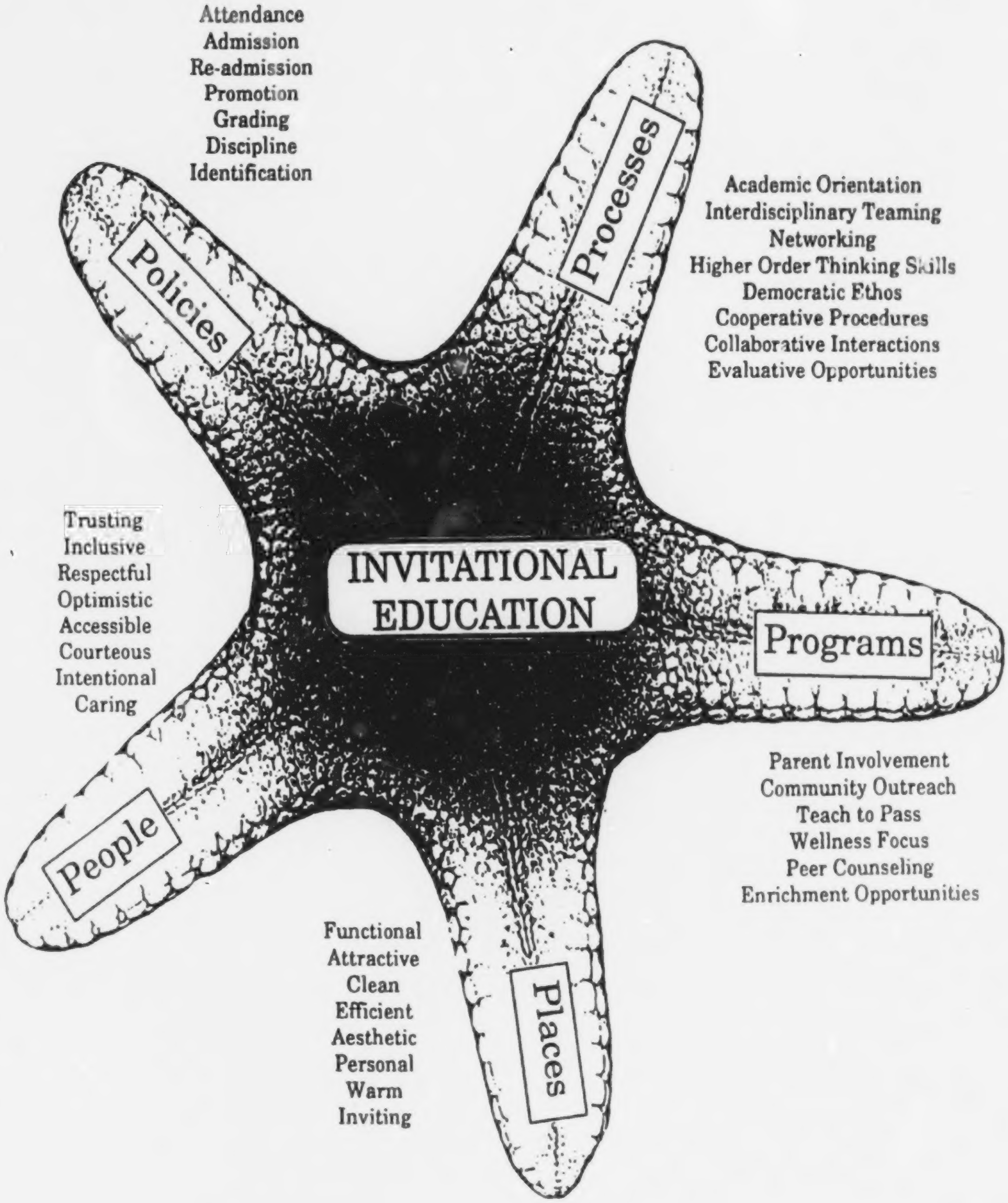

Attendance
Admission
Re-admission
Promotion
Grading
Discipline
Identification
Policies
Processes
Academic Orientation
Interdisciplinary Teaming
Networking
Higher Order Thinking Skills
Democratic Ethos
Cooperative Procedures
Collaborative Interactions
Evaluative Opportunities
Trusting
Inclusive
Respectful
Optimistic
Accessible
Courteous
Intentional
Caring
INVITATIONAL EDUCATION
Programs
Parent Involvement
Community Outreach
Teach to Pass
Wellness Focus
Peer Counseling
Enrichment Opportunities
People
Functional
Attractive
Clean
Efficient
Aesthetic
Personal
Warm
Inviting
Places

Figure 1. Starfish analogy illustrating the "Five-P" approach to school improvement.

Programs

In planning or revising *PROGRAMS*, we can be more innovative in finding ways to create more meaningful connections with our students, the curriculum, and the world around us. Sometimes, school programs, while well-meaning, can be elitist, discriminating, and counter-productive.

Processes

Finally, the very *PROCESSES* we employ to transform schools need to be democratically inviting in and of themselves. How we go about creating a more exciting, satisfying, and enriching school becomes as important as defining the inviting school we want to become.

These "five P's" provide a framework for considering school improvements in a holistic fashion. One of the most basic tenets of Invitational Education is that "everything is connected." Purkey and Strahan (1986) have summarized the importance of connectedness in "the Jello Principle"; school and everybody in it is like a big bowl of jello: if you touch it, the whole thing jiggles, everything is connected to everything else.

Thinking about PEOPLE, PLACES, POLICIES, PROGRAMS, and PROCESSES provides a strategy for systematic transformation of the whole school. Great ideas for curriculum development have a higher probability for success when they are connected to the "whole bowl of jello."

From Factory to Family

Once we begin to explore the "connectedness" of school improvement efforts, we can begin to develop collaborative strategies that encourage participation. As noted earlier, one of the basic assumptions of Invitational Education is that "education is a cooperative, collaborative activity where process is as important as product." How we reach our goals becomes as important as the goals themselves. Working together can draw us closer together and, as a result, the entire "feel" of the school becomes more inviting.

In planning our procedures, it is important to remember that school improvement efforts of the past have often been hampered by "factory model" operating procedures. Customary practices often include "defining problems," "breaking them down," developing "step-by-step" solutions, and testing these solutions with "productivity measures." The effect of such procedures is often isolation among committees and, sometimes, competitiveness among participants for "incentives."

Factory model procedures in schools rarely work. Not only is production a poor metaphor for schooling, there is a growing recognition among leaders in business and industry that traditional "assembly line thinking" does not work in the private sector either. In a speech at Duke University, W. Edwards Deming, the American "guru" of Japanese management principles, noted that "an every man for himself" atmosphere emphasizes quantity over quality. In contrast to production quotas, short term goals, incentive pay, and annual appraisals, Deming advocates a team concept that enlists the cooperation of workers to constantly improve processes and products. The success of Japanese firms that have taken his advice attests to the potential for Deming's participatory approach to management.

Our work with teachers and administrators from the seven schools collaborating with the Alliance for Invitational Education, headquartered at The University of North Carolina at Greensboro, has convinced us that teamwork is equally essential for schools. In contrast to a factory model orientation, Invitational Education advocates a "family model." Just as strong families grow stronger in working together for mutual support, successful schools grow more successful through teamwork. We have learned that the more "collegial" the relations among the teachers and administrators, the more dramatic the progress toward school transformation.

Our findings reinforce Barth's (1990) conclusions that the most powerful predictor of student achievement is the quality of relationships among the staff. His work with administrators and teachers over the past ten years through the Harvard Principals' Center has underscored the value of "shared leadership." Given the connectedness of all of the aspects of schooling, it is impossible for administrators and committee leaders to do everything that needs to be done. Moreover, when everyone participates in school transformation, everyone feels a genuine sense of ownership of the process. In fact, "membership" is a hallmark of an inviting school. The most inviting schools are those where everyone feels that he or she belongs.

The G.O.A.L.S. Process

The G.O.A.L.S. Process that we developed in our Alliance for Invitational Education projects provides a means for encouraging and sustaining membership in the school transformation process. The five basic steps of the process are designed to maximize involvement by everyone in the school: Goal setting, Outlining actions, Anticipating obstacles, Listing alternatives, and Specifying action plans.

The G.O.A.L.S. Process begins by organizing five "strand teams" (People, Places, Policies, Programs, and Processes) to begin thinking about ways to make the school more inviting. Strand teams work best when everyone is represented. Ideally, each strand includes administrators, teachers, staff members, volunteer parents, and volunteer students. Representatives from each of the five strands form a "Steering Committee" to oversee the process.

The G.O.A.L.S. Process provides a framework for identifying goals and implementing action plans in a collaborative fashion. Adapted from the "Pass it on exercise" developed by Doug Mac Iver (1991) at the Center for Research on Elementary and Middle Schools, John Hopkins University, the G.O.A.L.S. Process can be used within each strand to create an agenda for action and then used across strands to set priorities for implementation. Once each of the five strand teams identifies its goals, the list of goals is passed on to another strand team. That team reviews the goals and suggests actions. Goals and actions are then passed on to another team that anticipates obstacles. By generating a list of possible barriers, this step in the process provides a way for participants to explicitly address the type of "Yes, but . . ." statements of doubt that sometimes plague school improvement projects. Once obstacles have been identified, another team lists alternative strategies for implementation. The entire list of goals, actions, obstacles, and alternatives is then returned to the original strand team, which uses this feedback to develop an action plan.

The process provides a systematic strategy of collective planning and action. The process itself invites involvement. In using this process, we have found that participants feel empowered to make changes that reflect a consensus among their colleagues. We have discovered that each strand participating in the G.O.A.L.S. exercise finds that its original list of goals "comes back" enriched with many new ideas and suggestions. Resulting action plans are often more sophisticated than those developed independently. Not only have goals been identified and strategies suggested, but strategies are "tested" against anticipated obstacles and operationalized so that everyone can see how, when, and by whom they will become realities. More importantly, participants realize that school improvement is more than a "project," an effort with a beginning and ending date; it is a continuous force to enhance the entire climate of the school.

Invitational Education in Action

Invitational Education has been applied not only to the seven middle-level schools on which this article is based, but also to over one hundred schools throughout the United States and Canada. It will be useful here to take a closer look at some of the ways educators in one of the these schools, Douglas Byrd Junior High, Fayetteville, North Carolina, have put this theory into practice.

In the spring of 1990, Douglas Byrd Junior High School faced a dilemma. The faculty and staff were dedicated and hard-working, but they felt that they were losing the battle against many challenges they faced. The high dropout and absentee rates were the highest of the 12 junior high schools in Cumberland County, North Carolina, while the scores on standardized tests were the lowest. These and other challenges made Douglas Byrd an excellent testing ground for Invitational Education.

When Byrd Junior High was selected to become one of RJR Nabisco's *Next Century Schools*, teachers and administrators began meeting with the authors and other members of the Alliance for Invitational Education to plan a systematic approach to staff development and school improvement. Everyone involved in this planning agreed from the beginning that the intent was not merely to "reorganize" services or "restructure" education at Byrd, but rather to "transform" the quality of life in school for everyone involved, to deal with the "whole bowl of jello." The plan was to offer a unique "Opening of School Celebration" for everyone at Byrd, a leadership training program, and an ongoing series of inservice workshops. All of these activities were coordinated by a steering committee of teachers and administrators, and all were built around Invitational Education.

At the Opening of School Celebration, teachers, administrators, representative students and their parents gathered in a first-class environment at a Fayetteville hotel for a day-long celebration of promises and plans for the future. The location was selected to let participants experience first hand an inviting environment. After a general session that presented the basic concepts of Invitational Education, participants were divided into five "strand" sessions organized by the "five P's" of Invitational Education (People, Places, Policies, Programs, and Processes). Discussion in the strand sessions focused on three areas: current strengths, shared concerns, and suggestions for improvement. Strands began working through the

G.O.A.L.S. Process to maximize involvement: Goal setting, Outlining actions, Anticipating obstacles, Listing alternatives, and Specifying action plans. Strand leaders were identified to coordinate meetings throughout the semester and report to the steering committee.

During the first two years of the project, leadership training sessions provided opportunities to discuss and coordinate the G.O.A.L.S. Process. Action plans were circulated and priorities established across the strands. Inservice sessions were planned to facilitate strand goals. Task groups began meeting to address specific needs. Among the many workshops conducted were sessions on classroom discipline, conflict management, cooperative learning, teaming, and student evaluation. In all of these sessions, Invitational Education provided an organizing theory and a common language for school transformation.

As the project continued, results were dramatic. Major goals have included increasing community involvement, student recognition and faculty morale (People); renovating commons areas and developing comfortable spaces for students and faculty to congregate (Places); creating a more positive approach to discipline and ongoing academic support (Policies); increasing community involvement and encouraging students to stay in school (Programs); and developing strategies for cooperative learning and interdisciplinary teaming (Processes).

One of the most interesting findings was the relationship between Invitational Education and student self-concept-as-learner. *The Self-Concept-As-Learner Scale* (SCAL) (Purkey, Cage & Fahey, 1973) was administered to 175 students in the seventh grade and readministered to the same 175 students in the ninth grade. It was during this period that Invitational Education was introduced and implemented throughout the school. Results indicated the SCAL scores of the students remained stable over the two-year period. Their self-concept-as-learner scores did not decline as would have been expected on the basis of the findings of numerous previous studies that consistently report dramatic declines in student self-concepts through the school years (Stanley & Purkey, 1994).

Conclusions

Within the framework of Invitational Education, school transformation is seen as an ongoing process. While the outcomes of this process are measurable, the changes that matter most are often intangible. Our experiences with Invitational Education at these schools and others have helped us understand more about these intangibles. The changes that have mattered most are those that affect how teachers and students see themselves, each other, and their school. As students and teachers develop more positive views, the momentum of Invitational Education grows and the deep structures that shape the cultures of schools begin to change. Over time, schools become less like factories and more like families, where people want to spend their time and everyone is summoned cordially to realize their relatively boundless potential.

References

Barth, R. (1990). *Improving schools from within.* San Francisco, CA: Jossey-Bass Publishers.

Deal, T. (1990). Reframing reform. *Educational Leadership, 47*(8), 6-12.

Mac Iver, D. (1991, October). *The "Pass it on" exercise.* Presented at the Florida Regional Conference of the National Middle School Association. Fort Lauderdale.

Oberg, A. (1987). The ground of professional practice. In J. Lowyck (ed.), *Teacher thinking and professional action.* (Lisse: Swets & Zeitlinger).

Purkey, W. W., Cage, B. N., & Fahey, M. (1973). *The Florida key manual.* The University of North Carolina at Greensboro: Alliance for Invitational Education.

Purkey, W., & Novak, J. (1996). *Inviting school success: A self-concept approach to teaching, learning and Democratic practice,* (3rd ed.). Belmont, CA: Wadsworth Publishing Co.

Purkey, W., & Novak, J. (1988). *Education by invitation only.* Bloomington, IN: Phi Delta Kappa Education Foundation Fastback 268.

Purkey, W., & Strahan, D. (1986). *Inviting positive discipline.* Columbus, OH: National Middle School Association.

Stanley, P. H., & Purkey, W. W. (1994). Student self-concept-as-learner: Does Invitational Education make a difference? *Research in the Schools, 1*(2), 15-22.

Strahan, D. (1990). From seminars to lessons: A middle school language arts teacher's reflections on instructional improvement. *Journal of Curriculum Studies, 22,* 233-251.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 7-14

The Effect of Random Class Assignment on Elementary Students' Reading and Mathematics Achievement

Jayne B. Zaharias
Tennessee State University

C. M. Achilles
Eastern Michigan University

Van A. Cain
Tennessee State University

*The purpose of this study was to determine if random or non-random assignment to classes provides achievement benefits to students in Grades 1-3. Achievement measures were the Total Reading and Total Mathematics sub-scores from the Stanford Achievement Test (SAT) and the reading and mathematics sub-scores from the Basic Skills First Test (BSF). Data sources were the randomly assigned regular-size classes (22-26 students per class) from the longitudinal, experimental Student/Teacher Achievement Ratio (STAR) study and non-randomly assigned classes in comparison schools from STAR districts. Analyses included cross-sectional comparisons of randomly-assigned students (**n**=499) and non-randomly assigned students (**n**=658) using class means (between 18 and 28 classes per condition) in Grades 1, 2, and 3 using ANCOVA. Results favored random assignment of students by Grade 3 (p≤.01) as measured by SAT and BSF reading and mathematics sub-scores.*

Introduction

Although it is not possible to have a completely homogeneous class grouping (just having males and females adds heterogeneity), teachers and administrators have typically tried to reduce variability by assigning students to classes based on some indicator (e.g., reading scores) or by utilizing within-class instructional units. The present study compared test scores of students assigned at random with the test scores of students assigned by "usual," but not random, assignment procedures. The logical area of interest was to determine if random student assignment (K-3) resulted in different achievement levels from the usual class-assignment methods.

The Homogeneous vs. Heterogeneous Class Assignment Debate

Researchers began to study the effects of homogeneous groupings on academic achievement as early as the 1920s, and the outcomes of the vast amount of research have been debated off and on in the education literature for years. With the current interest on multicultural education and equity in education, the issue of homogeneous versus heterogeneous groupings has again emerged as an important consideration for educators.

Some recent literature focuses on the social implications of homogeneous grouping. One line of research supports beliefs that homogeneous grouping creates ethnic segregation within schools and inequitable learning opportunities for a large number of students (Braddock & Slavin, 1992; Oakes, 1990; Wheelock & Hawley, 1992). Because of the intention of grouping strategies to cater to students of differing abilities, students who are labeled as the "high" achievers or the "bright" group tend to be exposed to lower class sizes, more successful teachers, higher expectations, and a more enriched curriculum.

Jayne Boyd Zaharias is Research Director and Van A. Cain is Research Associate of the School-Community Partnership Unit within the Center for Research in Basic Skills at Tennessee State University. C. M. Achilles was a Project STAR Principal Investigator who is currently Professor of Educational Leadership at Eastern Michigan University. The authors thank Paul Madden, Professor of Educational Administration at Tennessee State University, for his direction of the doctoral dissertation on which this study was based. Authors are especially grateful to Helen Pate-Bain and Elizabeth R. Word for their leadership roles in Project STAR. Recognition is extended to Barbara A. Nye, Executive Director at The Center for Research in Basic Skills, Tennessee State University, for her support of this study. Special thanks to B. DeWayne Fulton and Dana A. Tollett for their input to this study. Requests for reprints should be sent to Jayne Boyd Zaharias, The Center for Research in Basic Skills, Tennessee State University, 330 10th Avenue North, Suite J/Box 141, Nashville, TN 37203-3401.

The "low" achievers or the "slow" students all too often experience just the opposite (George, 1992). As these inequities perpetuate so does the gap between high and low ability students' test scores and even their educational and occupational aspirations. Kozol (in Scherer, 1992/1993) elaborates:

> The little girl who gets shoved into the low reading group in 2nd grade is very likely to be the child who is urged to take cosmetology instead of algebra in the 8th grade, and most likely to be in vocational courses, not college courses, in the 10th grade, if she hasn't dropped out by then. So, it's cruelly predictive. (p.8)

Even with evidence of negative social effects, educators continue to utilize ability-grouping practices partly because many practitioners strongly believe that this is the best way to help students succeed academically. The position is not generally supported by research. Findley and Bryan's (1971) comprehensive review of the research on ability grouping indicates that homogeneous grouping of students has little or no value for raising academic achievement, nor did such methods expose students to more effective learning environments. Some studies have shown slight achievement gains favoring those students who were placed in high groups. These slight gains are "more than offset by evidence of unfavorable effects on the learning of students of average and below average ability, particularly the latter" (Esposito, 1973, p.171).

Slavin's (1987, 1990) reviews of the research on ability grouping show that whereas ability grouping at the secondary school level is not effective (and in some cases it negatively affected student achievement), certain types of ability grouping can have positive effects on student achievement at the elementary school level. The Joplin Plan, which regroups students across grades for reading instruction (Effect Size or ES = +.44), and within class ability grouping used for mathematics instruction in the upper elementary grades (ES = +.34) both appear to be effective grouping methods. There was not conclusive evidence to suggest either positive or negative effects for regrouping across same-grade classes in reading and mathematics. The synthesis of ability-grouped class assignment revealed an effect size of zero indicating that this practice does not improve student achievement.

Ability-grouped class assignment, one of the more common types of homogeneous grouping, ironically is the method that has received the least support from the research and has not been thoroughly investigated (Slavin, 1987). Slavin identified only 14 studies of ability-grouped class assignment that met the *a priori* criteria to be included in his research synthesis of the effects of ability grouping on elementary students' achievement.

The present study contributes to previous research on ability-grouped class assignment by employing a large extant randomized database to determine if random assignment (believed to result in heterogeneity) to self-contained classes versus "other" (non-random, presumably more homogeneous) types of class assignment has an effect on the academic performance of elementary students. If particular grouping strategies are shown to be effective or ineffective, this would provide educators and education administrators with another option for restructuring that would be relatively easy to implement at very little cost (Slavin, 1987).

While homogeneous grouping may have (arguably) helped teachers in the past, recent advances in learning theory, instructional modalities (e.g., peer tutoring, cross-age/grade tutoring, "hands-on" methods, learning centers, student learning teams, cooperative learning) and developmental understandings of early learners have provided vastly improved armamentaria for teachers. Coupled with the changing demographics of pupils entering schools today (e.g., Hamburg, 1992; Hodgkinson, 1992) and concerns that homogeneous classroom assignments have detrimental social effects on students assigned to low-track classes (which may be a form of "tracking" outlawed in Hobson v. Hansen, 1967, 1971 as cited in Reutter, 1985), information on the achievement differences of random and non-random class assignments will be useful in organization issues of school restructuring.

Scope of the Database:
Background on Project STAR

Cooley and Bickel (1986) emphasize the use of existing databases to investigate new policy questions as they arise. Utilizing extant data is less intrusive to school practitioners and shortens the time required to provide information. Although the STAR study was completed in 1990, the unique database lends itself to a wide array of educational research and continues to be used for subsidiary studies such as this recently completed study of random class assignment.

During 1985 the Tennessee legislature authorized and funded a major longitudinal experimental study, the Student/Teacher Achievement Ratio (STAR) Project, to determine the effects of class size on elementary (K-3) students' academic achievement. This study has contributed to class-size research by providing educators with definitive answers to many questions concerning the benefits of class size (Finn & Achilles, 1990).

The significance of the findings from STAR are important because of the study's unusually sound methodological design and its comprehensive, randomized database. As a reactor at the annual meeting of the American Educational Research Association, Slavin praised STAR's design and integrity and called it a "watershed event" in research. The STAR database contains many demographic and academic achievement variables on over 10,000 elementary students from as many as 79 Tennessee schools, and information from instruments and questionnaires completed by the principals, teachers, and teacher aides who participated in the study.

A consortium of four universities was formed to conduct the study which was directed by the Tennessee Department of Education. The Tennessee legislation (House Bill 544) established specifications for selecting schools to participate in the study: (a) enough schools from different locations (inner-city, rural, urban, suburban) had to be included in the study to compare the effects of class size by school type; (b) schools had to plan to remain in the study for four years; (c) schools had to agree to the use of random assignment of teachers and students; (d) schools had to have enough students (i.e., 57) at the kindergarten level (1985) to meet the study's within-school design (at least one small class with a range of 13-17 and one regular and one regular-with-aide class with each having a range of 22-26 students); and (e) all teachers had to be certified at the appropriate grade level. The final participants included 79 schools in 42 systems which resulted in the first-year (kindergarten) sample of over 6,000 students (Word et al., 1990).

For each grade level (K-3) standardized achievement test scores and demographic variables (i.e., sex, race, birthdate, socioeconomic status) have been retained in the database. The STAR database also contains data on more than 1,000 students from 21 comparison schools located in 13 STAR districts. The STAR principal investigators determined that these schools had operational characteristics similar to the STAR schools. During each year of the study, comparison school students were administered the same achievement tests as the STAR students, and the same demographic data were collected on the comparison students (Word et al., 1990). However, the extensive classroom data collection (e.g., teacher profiles, within-class grouping survey, teacher problem checklist) that was conducted in STAR schools was not replicated in the comparison schools.

At the beginning of the STAR study (1985-86 school year), kindergarten students were randomly assigned to one of three class-size conditions: small (13-17 with a mean of 15), regular (22-26 with a mean of 24), or regular with a full-time teacher aide. Students who entered the schools during the study were randomly assigned to one of the three conditions. Teachers were randomly assigned to classes at the beginning of each school year. *The random assignment and the class-size changes were the only modifications to the usual practices of the STAR schools, and there were no interventions in the comparison schools.*

Design and Analysis of the Present Study

For the purposes of this study a sub-sample of students from only those school districts that contained both project schools and comparison schools was drawn from the STAR database. The present study essentially isolated the variable of random assignment at the class level as the treatment and some non-random form of class assignment as the control situation. The experimental group for this study was selected from the pool of STAR students who were in *regular-size classes*. Therefore, random assignment was the only special treatment to this group.

To determine what types of class-assignment procedures were used in the comparison schools, telephone interviews were conducted with school administrators and/or faculty who were employed at those schools during 1985-1989. Sixteen of the 21 comparison schools used some form of non-random assignment such as: previous teacher's recommendation, current teacher's recommendations, reading test scores, mathematics test scores, retention, discipline, gender, ethnicity, or a combination of these variables to achieve "a good mixture" of students in each class. Five schools were reported to have used random assignment and therefore were excluded from the statistical analyses. Since various assignment methods were used throughout the comparison schools, data could not be aggregated into subsamples of sufficient size to compare each different assignment method to random assignment. Both STAR regular classes and the comparison-school classes used within-class groupings for reading and mathematics. Analyses in this study compared the effects of random-class assignment (STAR/experimental schools) to other types of class assignment (comparison schools in the same districts as STAR schools) on achievement outcomes.

Since a major finding of STAR was a positive class-size effect on pupil achievement, class-size ranges and mean class sizes were calculated for students at each grade level in the random and non-random groups. Out-of-range classes, that is classes containing too many or too few students (under 16 or over 30), were dropped from statistical analyses. An ANOVA showed that there were no significant differences ($p < .05$) between the mean class sizes of the random and non-random groups (see Table 1).

Table 1
Class-Size Means and Ranges by Condition
(Grades 1-3) from a Study of Random Class Assignment,
Project STAR Database (TN, 1985-1989)

	Random		Non-Random	
Grade	*M*	Range	*M*	Range
1	22.50	20-30	23.43	16-28
2	23.30	19-29	22.26	17-26
3	19.32	16-23	20.53	16-25

A chi-square was applied to determine if there were statistically significant differences between the percentage of males and females and the percentages of whites and non-whites between the experimental and comparison schools. Four comparison schools had significantly more minorities than the experimental schools and were excluded. Tables 2 and 3 show the resulting gender and ethnic ratios.

Table 2
Number and Percentage of Students by Gender
(Grades 1-3) from a Study of Random Class Assignment,
Project STAR Database (TN, 1985-1989)

Grade	*n*	Female	Male
1			
Random	667	323 (48%)	344 (52%)
Non-Random	657	302 (46%)	355 (54%)
2			
Random	625	297 (48%)	328 (52%)
Non-Random	576	278 (48%)	298 (52%)
3			
Random	613	291 (47%)	322 (53%)
Non-Random	314	158 (50%)	156 (50%)

Table 3
Number and Percentage of Students by Ethnicity
(Grades 1-3) from a Study of Random Class Assignment,
Project STAR Database (TN, 1985-1989)

Grade	*n*	Minority	White
1			
Random	667	50 (7%)	617 (93%)
Non-Random	656	29 (4%)	627 (96%)
2			
Random	625	48 (8%)	577 (92%)
Non-Random	576	18 (3%)	558 (97%)
3			
Random	613	43 (7%)	570 (93%)
Non-Random	314	14 (4%)	300 (96%)

For each comparison school that was eliminated from analyses, due either to reported use of random assignment or statistically significant differences in ethnicity, the experimental schools located within the same school systems were also excluded. This procedure was followed to keep the two groups as comparable as possible in regard to system characteristics (e.g., per pupil expenditure, district policies and procedures). After all exclusions were made, student data from 16 experimental schools and 12 comparison schools were available for analyses. According to STAR research guidelines, two of these schools (one experimental and one comparison) were classified as urban. All other schools included in this study were considered rural.

The total number of students available for analyses was 1,157. Random (n = 499) and non-random group (n = 658) sample sizes vary between grade levels. Sample sizes also fluctuate between the analyses of reading and mathematics as the database contains students who had test scores for only one of the subject areas. Those students who remained in the STAR study during first, second, and third grades; first and second grades; and second and third grades were identified for analyses of cumulative effects. The n's for analyses of cumulative effects also vary slightly by grade and because a student may not have completed all tests.

Outcome measures were the Stanford Achievement Test (SAT) Total Reading and Total Mathematics scaled scores and the Basic Skills First (BSF) Total Reading and Total Mathematics raw scores (Grades 1-3). The SAT is a norm-referenced test (NRT) that measures student achievement in reading/language arts, mathematics, science, and social science based on national norms (Gardner, Madden, Rudman et al., 1983). The BSF is a criterion-referenced test (CRT) developed by the Tennessee Department of Education to measure mastery of the state curriculum in Grades 1, 2, and 3 (Tennessee Comprehensive Curriculum Objectives, 1989).

To compare the mean reading and mathematics test scores between random and non-random groups for statistically significant differences, one-way ANCOVAs controlling for ethnicity and gender were applied to student-level and class-aggregate cross-sectional data sets (Grades 1, 2, 3).[1] Student-level data were employed for execution of ANCOVAs controlling for previous test scores, ethnicity, and gender to determine statistically significant cumulative effects (two or more years of treatment). For each analysis the alpha level was set at $p \leq .05$.

Findings[2]

In reading, the randomly-assigned students outscored the non-randomly assigned students on both the SAT and BSF measures at each grade level. At Grade 2 the difference in SAT means was only slightly larger than at Grade 1 (approximately 3 versus 4 scaled scores) and there were practically no differences (less than one raw score) at Grades 1 and 2 on the BSF. Differences at Grade 3 reached statistical significance on both the SAT ($p \leq .05$) and BSF ($p \leq .01$) achievement measures (see Table 4).

Table 4
Cross-sectional Effects of Random Class Assignment on Reading Achievement as Measured by Class-aggregate SAT Total Reading Scaled Scores and BSF Total Reading Raw Scores, Project STAR Database (TN, 1985-1989)

	Random		Non-Random				
Grade	*n*	*M*	*n*	*M*	S.D.	F	E.S.
				SAT			
1	21	532.95	23	529.35	29.52	.209	.12
2	19	596.95	22	592.86	17.81	.618	.23
3	15	625.61*	19	615.56	11.36	4.291	.88
				BSF			
1	22	27.22	23	27.13	2.51	.075	.04
2	19	40.85	22	40.75	3.60	.028	.03
3	17	33.94*	14	32.12	1.52	6.177	1.20

*$p \leq .05$.

Analyses of the mathematics outcomes, although not significant, favored the non-random group at Grade 1 on both the SAT and BSF measures. Table 5 shows that at Grade 2 the experimental group was scoring somewhat higher on both measures, and by the end of Grade 3 the differences were statistically significant in favor of the randomly-assigned students (SAT: $p \leq .001$; BSF: $p \leq .01$).

Table 5
Cross-sectional Effects of Random Class Assignment on Mathematics Achievement as Measured by Class-aggregate SAT Total Math Scaled Scores and BSF Total Math Raw Scores, Project STAR Database (TN, 1985-1989)

	Random		Non-Random				
Grade	*n*	*M*	*n*	*M*	S.D.	F	E.S.
				SAT			
1	22	540.64	23	545.87	24.76	.427	-.21
2	19	590.95	22	587.18	17.81	.460	.28
3	19	631.95*	18	616.67	11.36	7.981	1.43
				BSF			
1	22	40.25	23	40.71	2.51	.527	-.24
2	19	53.87	22	53.48	3.60	.232	.15
3	18	53.65*	13	50.30	1.52	8.017	1.38

*$p \leq .05$.

To determine if any differences between random assignment and other methods of class assignment on students' cumulative reading and mathematics achievement were significant, ANCOVAs controlling for ethnicity, gender, and previous test scores were applied to the mean SAT Total Reading and Total Mathematics gain scores of students who remained in the study from Grade 1 to Grade 2, Grade 2 to Grade 3, and Grade 1 to Grade 3. Table 6 shows that the only significant difference in reading gain scores occurred from Grade 2 to Grade 3 and favored the random group ($p \leq .01$). The randomly-assigned group made greater mathematics gains than the control group in all three comparisons, but the mean differences were significant for only the Grade 2 to Grade 3 ($p \leq .001$) and Grade 1 to Grade 3 ($p \leq .001$) analyses.

Table 6
Cumulative Effects of Reading and Mathematics Achievement for Randomly Assigned (Experimental) and Non-randomly Assigned (Control) Students as Measured by Individual-level SAT Total Reading and Math Gain Scores, Project STAR Database (TN, 1985-1989)

	Random		Non-Random				
Grades	*n*	*M*	*n*	*M*	S.D.	F	E.S.
				Reading			
1-2	328	59.74	465	63.87	43.07	1.7450	-.10
2-3	311	21.71*	425	20.54	37.29	.0521	.03
1-3	261	81.77	324	78.48	38.71	.0679	.08
				Mathematics			
1-2	331	43.63	465	40.34	43.71	2.5077	.08
2-3	311	36.83*	426	24.51	35.11	14.1366	.35
1-3	264	81.88*	324	63.48	30.89	53.5634	.60*

$p \leq .05$.

To summarize the findings, out of the 18 analyses (9 comparing reading scores and 9 comparing mathematics scores) 15 favored the randomly-assigned students but only 7 were significant (3 in reading and 4 in mathematics). Three out of the 18 analyses favored the non-random group but none were significant. A signs test determined that the number of differences between groups (whether statistically significant or not) favored ($p \leq .01$) the randomly-assigned group.

Summary and Discussion

Based on the findings of this study, random assignment to classes appears to increase the reading and mathematics achievement of early elementary students (K-3). In most cases, the baseline scores of the random group (Grade 1) were generally either equal to or lower than the scores of the control group. At Grade 2 the scores of the random group began to surpass those of the control group. By Grade 3 most scores of the random group were higher than the control-group scores. The trend of positive effects from random class assignment (heterogeneity) on mathematics achievement was fairly clear. The effects of random class assignment on reading achievement were less evident.

The inconsistency of the effects of random assignment on reading achievement might be attributed to the pedagogical tendency to form within-class reading ability groups. That is, even though teachers were presumably assigned a heterogeneous group of students they formed several homogeneous groups within the classroom for the purpose of teaching reading. This practice may have diluted the possible benefits of class-assigned heterogeneity. Since data on within-class groupings were collected only for the STAR schools and not the comparison schools, this study could not control for effects of within-class grouping. Therefore, it is necessary to extrapolate some information from the literature and analyses of the STAR study to support this conclusion.

The literature reports that within-class ability grouping is the most common form of homogeneous grouping. This practice is so common, in fact, that few studies of the effects of within-class ability grouping on reading achievement have been conducted due to the difficulty in arranging for ungrouped control groups even on a temporary basis (Slavin, 1987). Data from the Project STAR grouping questionnaire support that the randomly-assigned classes were divided into several (usually three) within-class reading groups based on the teachers' evaluation of students' ability levels. Data available for all classes used in the aggregate analysis of effects on reading showed that all of these classes (100%) used within-class reading groups based on skill level. At Grade 1 students were divided into from two to five groups; Grade 2 classes contained two to four reading groups; and at Grade 3 from two to three reading groups were formed within classes. The decline in the number of groups at Grade 3 (50% of the teachers reported using only two groups and 50% reported using three groups) may help explain the onset of gains in the random group's reading scores at Grade 3.

In contrast, the STAR grouping questionnaire data showed that only 13% of the first-grade, 21% of the second-grade, and 26% of the third-grade classes included in the present study were grouped for mathematics. Only one third-grade teacher reported using three groups; teachers in all other randomly assigned classes reported using only two groups. Fewer groups should provide a more heterogeneous atmosphere thereby enhancing the opportunity for any possible positive effects from random class assignment to emerge.

This study demonstrates positive effects on reading and mathematics achievement of students who were randomly assigned to classes. Given this finding, the previous research supporting the results, and the current literature indicating that homogeneous groupings perpetuate detrimental social effects (e.g., Glickman, 1991; Oakes, 1992; O'Neil, 1993; Wheelock, 1992), then why is random class assignment not being used by more schools? One plausible reason is that such practices have been embedded in pedagogy. It appears that only within the last several years have researchers identified successful strategies for the teaching of heterogeneous groups. Moreover, it has been very recently that colleges and universities have begun to prepare teachers with the new methods that are considered viable alternatives to ability grouping.

A second reason stems from the fact that school administrators, teachers, parents, and the general public have become too comfortable with the current system. The "this is the way we've always done it" mentality serves as the glue which continues to keep old norms and beliefs stuck in the minds of most people. The change process is never easy. Nevertheless, to insure that all children are equitably provided with the best learning opportunities, the system has to change. Random class assignment is an inexpensive and relatively easy to implement restructuring method that maximizes the academic achievement of children. By using random assignment, teachers and administrators could spend less time deciding how to sort children into classes. The valuable time that the usual class assignment efforts consume could be used more effectively for other planning activities, and a computer could randomly sort students in a matter of minutes.

References

Braddock, J. H., & Slavin, R. E. (1992, September). *Why ability grouping must end: Achieving excellence and equity in American education.* Paper presented at the Common Destiny Conference, Washington, D.C.

Cooley, W., & Bickel, W. (1986). *Decision-oriented educational research.* Boston: Kluwer-Nijhoff Publishing.

Esposito, D. (1973). Homogeneous and heterogeneous ability grouping: Principal findings and implications for evaluating and designing more effective educational environments. *Review of Educational Research, 43*(2), 163-179.

Findley, W. G., & Bryan, M. (1971). *Ability grouping: 1970 Status, impact, and alternatives.* Athens: Center for Educational Improvement, University of Georgia. (ERIC Document Reproduction Service No. ED 060-595).

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*, 557-77.

Gardner, E. F., Madden, R., Rudman, H. C., Karlsen, B., Merwin, J. C., Callis, R., & Collins, C. S. (1983). *Stanford achievement test series: Group norms booklet.* The Psychological Corporation, Harcourt Brace Jovanovich Publishers.

George, P. (1992). *How to untrack your school.* Alexandria, Virginia: Association for Supervision and Curriculum Development.

Glickman, C. (1991, May). Pretending not to know what we know. *Educational Leadership, 48*(8), 4-10.

Hamburg, D. (1992). *Today's Children.* New York: Time Books, Random House.

Hodgkinson, H. (1992, June). *A demographic look at tomorrow.* Washington, D.C.: Institute for Educational Leadership.

Oakes, J. (1992). Can tracking research inform practice? Technical, normative, and political considerations. *Educational Researcher, 21*(4), 12-21.

Oakes, J. (1990, August). *Beyond tracking: Making the best of schools.* Cocking lecture given at the annual meeting of the National College Professors of Educational Administration, Los Angeles.

O'Neil, J. (1993, June). Can separate be equal? Educators debate merits, pitfalls of tracking. *ASCD Update*, 1-8.

Reutter, E. E., Jr. (1985). *The law of public education.* Mineola, New York: The Foundation Press.

Scherer, M. (1992/1993). On savage inequalities: A conversation with Jonathan Kozol. *Educational Leadership, 50*(4), 4-9.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*(3), 471-499.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57*(3), 293-336.

Slavin, R. E., & Karweit, N. L. (1985). Effects of whole class, ability grouped, and individualized instruction on mathematics achievement. *American Educational Research Journal, 22*(3), 351-367.

Tennessee Department of Education (1989). *Tennessee comprehensive curriculum objectives.* Nashville, Tennessee.

Wheelock, A. (1992). *Crossing the tracks: How "untracking" can save America's schools.* New York: The New Press.

Wheelock, A., & Hawley, W. D. (1992, September). *What next? Promoting alternatives to ability grouping.* Paper presented at the Common Destiny Conference, Washington, D.C.

Word, E. R., Johnston, J., Pate-Bain, H., Fulton, B. D., Zaharias, J. B., Lintz, M. N., Achilles, C. M., Folger, J., & Breda, C. (1990). *The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report*. Nashville, Tennessee: Tennessee State Department of Education.

Footnotes

[1] Analyses were initially conducted on both student-level means and class means to alleviate concerns that large sample sizes from individual means could result in inflated p values, *and* that class means could inflate effect sizes. The two cross-sectional data sets essentially showed the same results.

[2] Slavin (1987) notes that effect sizes using class-aggregate means and gain scores from post-test only data may be inflated. Therefore, such effect sizes reported for this study should be regarded as approximate indicators of treatment effects.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 15-21

Retention Across Elementary Schools in a Midwestern School District

Sim Gurewitz
Crete Public Schools

Jack Kramer
University of Nebraska - Lincoln

Rates of retention over a five-year period were analyzed by grade, by school, and by the socioeconomic status of each of 32 elementary schools in a Midwestern school district. Analysis of variance revealed that schools of Middle socioeconomic status retained students at a significantly higher rate than did schools of Low and High socioeconomic status. These findings contradict those of earlier studies, which found higher retention rates in schools of Low socioeconomic status. Differences among mean rates of retention between schools were highly variable; the highest retaining school retained at a mean rate over 100 times that of the lowest retaining school. Results suggest that the probability that a student will be retained is to a considerable extent a function of the school attended, rather than a function of individual student characteristics.

Retention, the practice of having a student repeat a grade, is widely employed and has been studied intensively. The primary goal of retention is to remedy academic inadequacy (Byrnes, 1989; Jackson, 1975; Tomchin & Impara, 1992). Research into this treatment for academic deficiencies has focused to a large extent on individual student characteristics (e.g., individual students' levels of academic achievement prior to retention) and outcomes (e.g., drop-out rate, level of academic achievement after retention). This body of research has not been supportive of the practice of having students repeat a grade. Holmes (1989), for example, performed a meta-analysis of sixty-three studies which employed comparison groups. This analysis found that on the outcome measures of academic achievement, personal adjustment, self-concept, attitude toward school, and attendance, the non-promoted pupils scored -.15 standard deviation units (weighted by effect; -.26 weighted by study) lower than the promoted comparison groups. The mean effect size for academic achievement alone was -.19, weighted by effect. This translates to a drop of three standard score points on a test of achievement with a standard deviation of 15, indicating that grade retention is not an effective academic intervention.

Despite these findings, retention affects substantial numbers of children. Although national data regarding the incidence of retention is not available, estimates indicate that retention is experienced annually by over one million students (Jackson, 1975; Smith & Shepard, 1987). In addition, it has been suggested that the rate of retention may be increasing in response to the minimum competency testing movement, national concerns regarding the effectiveness of the nation's schools (Rose, Medway, Cantrell, & Marus, 1983), and the educational reform/effective schools movement (Shepard & Smith, 1989).

The relationship of retention and the socioeconomic status (SES) of individual students has been analyzed extensively, with retention rates reported to be highest for students of low socioeconomic status (Abidin, Golladay, & Howerton, 1971; Bossing & Brien, 1980; Hersh, 1988; Mantzicopoulos, Morrison, Hinshaw & Carte, 1989; Safer, 1986). Given that retention is viewed primarily as an academic intervention, these results are consistent with the robust correlations generally found between socioeconomic status and academic achievement.

Fewer studies have considered the socioeconomic status of the school or the school district, as opposed to the socioeconomic status of individual students. The results of this body of research are generally consistent with the analyses concerning individual students, with lower socioeconomic status associated with higher retention rates. Safer, Heaton and Allen (1977) found that the retention rate in the blue collar area of a school

---

Sim Gurewitz is a School Psychologist for the Crete (NE) Public Schools and pursues his research interests in association with the University of Nebraska--Lincoln. Jack Kramer is in private practice after working in university settings for 15 years. He provides psychological services to schools, children, and families. Correspondence regarding this article should be directed to Dr. Sim Gurewitz, Crete Public Schools, 920 Linden Street, Crete, Nebraska 68333.

district was three times the rate of the white collar area of the same district. More striking was the rate at which students were retained more than once; on this measure the rate in the blue collar area was six times that of the white collar area. Similar findings were reported in Gastright's (1989) analysis of retention in 33 of the 43 member districts of the Council of the Great City Schools. Although wide variation in practices was reported, retention rates generally increased with declining socioeconomic status; the rate of the lowest SES group was twice the rate of the highest SES group. These findings are consistent with Morris' (1991) analysis of retention in the Dade County (FL) elementary schools over a five-year period. Here the High SES schools had the lowest retention rates, and Low SES schools had the highest rates, with Middle SES schools falling between the High and Low groups.

Two studies of retention across schools did not find a relationship between SES and retention rate. Hess and Greer (1987), in an analysis of the elementary schools ($N = 381$) in Chicago did not find significant differences in the use of retention based on the proportion of low income students enrolled. Similarly, Smith and Shepard's (1987) analysis of retention in Boulder, CO, found that high-retaining and low-retaining schools were not distinguishable according to socioeconomic status.

The purpose of this study was to review patterns of retention across schools in a Midwestern school district and to analyze the relationship between the schools' rates of retention and the schools' socioeconomic status.

Method

Annual data were obtained for each of 32 elementary schools in a Midwestern school district, Kindergarten through sixth grade, for a five-year period. The schools were ranked by socioeconomic status, and analysis of variance was used to explore differences in retention rates among schools of different socioeconomic status, to compare retention rates in different grades, and to evaluate overall trends in retention rate.

Subjects

The subjects of this study were all elementary schools ($N = 32$) in a Midwestern school district for the academic years 1984-1985 through 1988-1989. One school which contained only Kindergarten (K) through third grade was excluded from the analysis. The elementary school population during the period averaged 13,901 students per year; individual school populations ranged from 143 to 737 students ($\bar{X} = 434$). The data obtained for each of the five years included the total population of each of the grades from K through 6 in each school, and the number of retentions in each grade for each school.

Procedures

A panel of five raters was assigned to rank the schools from highest to lowest SES. The use of census tract data was rejected because school catchment area boundaries in the district are not contiguous with the census tracts, with many schools drawing students from three or more census tracts. The use of free/reduced lunch count per school was rejected, because although these data could have identified schools of Low SES, it would not have been possible to distinguish between schools of High and Middle SES. Therefore a panel of five raters with knowledge of the district was employed.

The raters were all employees (a media specialist, two special education teachers, a student services coordinator, and a Kindergarten teacher) of the district. Their tenure with the district ranged from 3 to 15 years ($\bar{X} = 10.4$ years). The raters were mailed ranking forms, with instructions to list the schools in rank order, from the highest to the lowest perceived socioeconomic status. For each list, we assigned the school of highest SES a rating of 1, and so forth, with the school of lowest SES assigned a rating of 32.

Inter-rater reliability between raters 1, 2, 4, and 5 ranged from .82 to .92 (Spearman correlation coefficients). Agreement of the other raters with rater 3 ranged from -.17 to .05. In light of the high level of agreement among raters 1, 2, 4, and 5, rater 3 was excluded from the analysis. The Spearman correlation coefficients obtained for all inter-rater comparisons are shown in Table 1.

Table 1
Interrater reliability; ratings of socioeconomic status of elementary schools

Raters 1, 2, 4, 5	r	Rater 3	r
1 - 2	.83	2 - 3	-.17
2 - 4	.83	3 - 5	-.02
2 - 5	.84	1 - 3	.05
1 - 4	.84	3 - 4	.07
4 - 5	.91		
1 - 5	.92		

The schools were then placed into socioeconomic groups based on natural breaks in the data. This division avoided placing schools with very close rankings into groups of different socioeconomic status. Based on these natural breaks in the rankings, 13 schools were judged to be of High socioeconomic status, 11 schools to be of Middle socioeconomic status, and 8 schools to be of Low socioeconomic status.

Results

Retention and the Socioeconomic Status of the School

Schools of Middle SES retained students at a higher rate than did schools of High and Low SES. The mean annual retention rate of the schools of Middle socioeconomic status ($\bar{X}$ = 1.5% of the student population retained annually; S = 1.2) was over four times the rate of the schools of High socioeconomic status ($\bar{X}$ = .35%; S = .34); the retention rate of schools of Low socioeconomic status ($\bar{X}$ = .95%; S = .5) fell between that of the Middle and High groups. One-way analysis of variance (High, Middle and Low socioeconomic status by retention rate) revealed that these differences were significant [$F(2,29) = 6.76, p \leq .004$]. Post-hoc pairwise comparisons using Tukey tests did not reveal significant differences.

The relationship observed between retention practices and school socioeconomic status is illuminated by examining those schools which retain at rates much higher or lower than the district average. The relationship of these schools' retention rates and socioeconomic status is illustrated by Figure 1, which shows the percentage of schools retaining at Z scores (standard deviation units above or below the mean retention rate for all schools) above +.5 and below -.5, by school socioeconomic status. Among schools of High SES, 8% obtained Z scores above +.5, and 77% obtained Z scores below -.5. Among schools of Low SES, 25% obtained Z scores above +.5 and 25% obtained Z scores below -.5. However, 63% of the schools of Middle SES obtained retention rate Z scores in excess of +.5, and only 27% obtained Z scores below -.5.

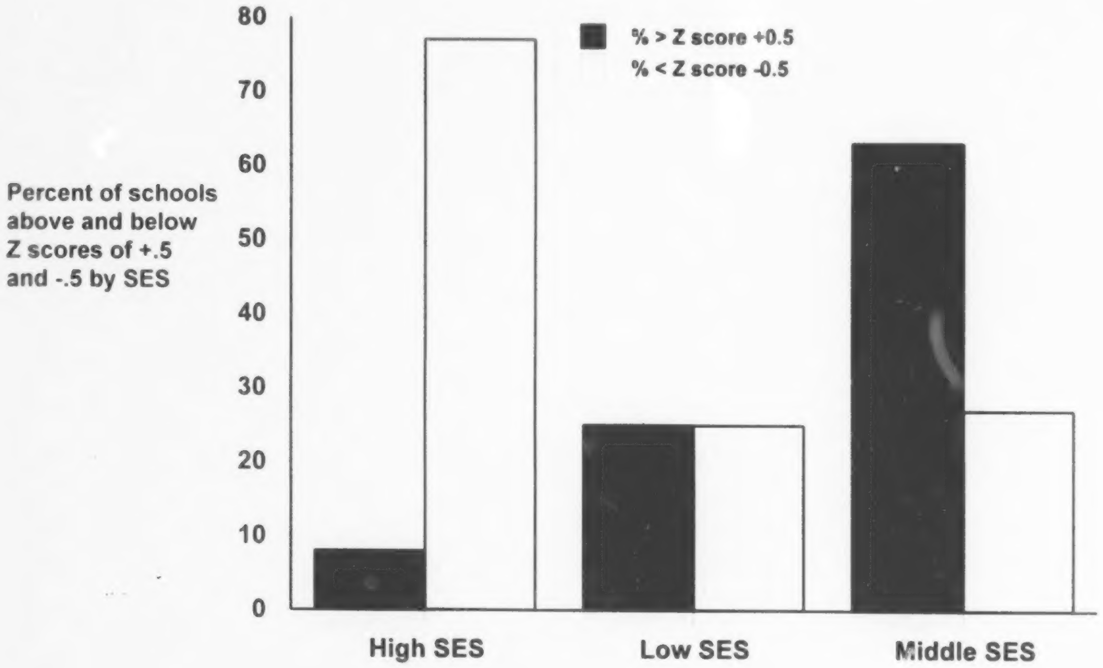


Figure 1. Percent of schools employing retention at rates above and below Z scores of +.5 and -.5, by socioeconomic status.

Retention Across Schools

A wide range of practices was observed within school socioeconomic groups. For example, within the group of High SES schools, the highest-retaining school retained at a rate 17 times that of the lowest-retaining school (.075% v. 1.24% retained per year). Within the group of Low SES schools, the highest-retaining school retained at a rate 30 times that of the lowest-retaining school (.05% v. 1.51%). Another perspective is gained by comparing the lowest and highest-retaining schools in the High, Middle, and Low SES groups. The lowest-retaining High, Middle, and Low SES schools retained at annual rates of .08%, .13%, and .05% respectively; among these three schools the highest-retaining school retained at a rate only 2.6 times that of the lowest-retaining school. The highest-retaining schools in the High, Middle, and Low SES groups retained at rates of 1.24%, 4.30%, and 1.51% respectively; among these three schools the highest-retaining school retained at a rate only 3.5 times that of the lowest retaining school (these relationships are illustrated in Figure 2).

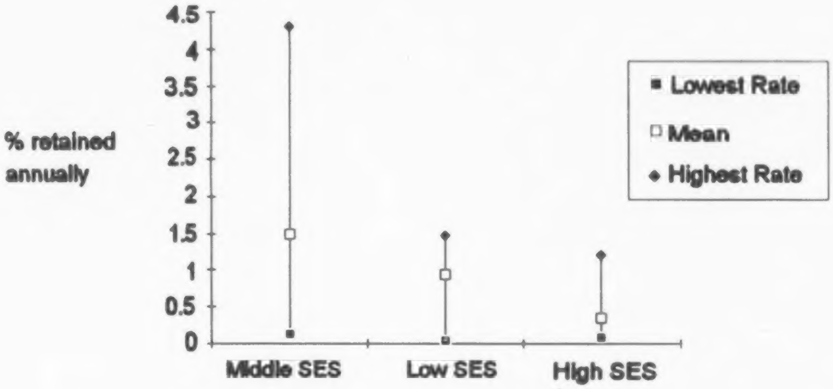


Figure 2. Range and Mean percent of students retained annually by school socioeconomic status.

The wide variation across schools can also be illustrated through comparison of specific grade level cohorts. In first grade, for example, there were no retentions in 93 of the 160 class cohorts (32 schools x 5 years of data). However, in each of two first grade cohorts, 20% of the students were retained. In Kindergarten, no students were retained in 86 of the 160 class cohorts, while 12.5% were retained during year 3 in school 3. The most extreme range was observed in third grade. Although no students were retained in 125 of the 160 grade cohorts, 20.7% of the students were retained during year 1 in school 3. Table 2 illustrates the mean percent retained, range, and standard deviation by grade during the five-year period.

Table 2
Range, Number Retained, and Mean Percent of Students Retained per Year by Grade Level

Grade	Range of % retained	Number of the 160 Cohorts Not Retaining	Mean % retained	Standard Deviation
KG	0 - 12.5	86	1.7	1.5
1	0 - 20.0	93	1.8	2.6
2	0 - 15.6	108	1.1	1.4
3	0 - 20.7	125	.77	1.4
4	0 - 7.1	142	.40	.71
5	0 - 6.3	147	.19	.33
6	0 - 2.8	152	.08	.17

Retention and Grade Level

Retention was employed most frequently in Kindergarten ($\bar{X}$ = 1.7%) and first grade ($\bar{X}$ = 1.8%). Thereafter the rate decreased steadily as grade level increased; by sixth grade the mean was .08% retained. Tests of difference among the means for the seven grades showed that grade level had a significant effect, [$F(6,186) = 12.33$, $p \leq .000$]. Tukey pairwise comparisons between grade levels confirmed that the retention rates for grades K, 1, 2, 3, and 4 differed significantly from the rates for grades 5 and 6.

Simple main effects tests were conducted to evaluate potential interaction among socioeconomic status, retention rate, and grade level. Significant effects were not found in fifth and sixth grades. Significant effects were found in Kindergarten [$F(2,29) = 5.12$, $p \leq .012$], first [$F(2,29) = 3.62$, $p \leq .040$], second [$F(2,29) = 14.91$, $p \leq .000$], third [$F(2,29) = 3.75$, $p \leq .036$], and fourth grades [$F(2,29) = 4.47$, $p \leq .020$]. Schools of Middle SES had the highest rate of retention for four of the five grades in which a significant socioeconomic status effect was noted. Figure 3 illustrates the relationship of retention rate, grade level, and socioeconomic status.

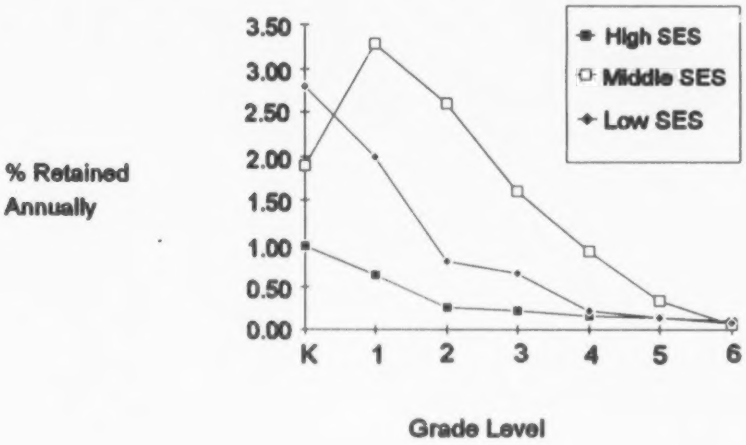


Figure 3. Retention rate by grade level and socioeconomic status.

Discussion

The strength and overall pattern of our data suggest that the socioeconomic status of the school is a significant factor in promotion/retention decisions. According to our data, the probability of a student being retained depends to a great degree on the student's school, rather than on the characteristics of the student. The differences in retention rates across schools are unlikely to be explained by proposing that the rate of student academic inadequacy is 33 times higher in school 15 than it is in school 4 (both schools of Middle SES). In addition, the relatively wide range of practices observed within socioeconomic groups suggests the operation of idiosyncratic factors in addition to a school socioeconomic factor. Thus it appears that although retention is an intervention implemented in response to an individual student's academic shortcomings, factors external to the individual student play an important and largely unexamined role in the decision-making process.

Our finding that schools of Middle SES experienced the highest rates of retention is at odds with Safer et al.'s (1977) analysis, and the analyses of Dade County, FL (Morris, 1991) and the member districts of the Council of Great City Schools (Gastright, 1989), all of which found the highest rates of retention among schools of lowest SES. Our findings also are not consistent with Smith and Shepard's (1987) analysis of Boulder, CO, and Hess and Greer's (1987) analysis of Chicago, IL; these studies found that school SES was not associated with retention rate.

Three scenarios which may explain a relationship between rate of retention and school socioeconomic status can be evaluated in light of our findings. The first explanation, that the retention rate would be highest in schools of Low SES, is consistent with research showing that students of low socioeconomic status are more likely to be retained (Abidin et al., 1971; Bossing & Brien, 1980; Hersh, 1988; Mantzicopoulos et al., 1989; Safer, 1986), and with the strong correlation between academic achievement and socioeconomic status. This line of reasoning thus fits well with the primary rationale underlying retention, which is that retention is an intervention employed to address academic inadequacy (Jackson, 1975). However, our findings suggest that factors other than the characteristics of the individual student can impact the probability that the student will be retained.

A second explanation is that the retention rate would be highest in schools of High socioeconomic status. The high rate could, paradoxically, be a function of the higher average level of achievement in schools of High socioeconomic status. In this environment, the low achieving student would be more conspicuous, and teachers' concerns regarding the low achiever's ability to keep up (Smith, 1989) might be greatest. Teachers also desire to reduce heterogeneity within the classroom (Smith, 1989). Therefore, this line of reasoning suggests that the low achieving student would be more likely to be retained in a school of High socioeconomic status. Additional factors arguing for a higher rate of retention in these schools are that parents in general are strong supporters of retention (Gallup, 1986; Gallup, 1983), and parents of higher socioeconomic status have more involvement with the school (Byrnes & Yamamoto, 1986) and higher academic expectations for their children. These parents might therefore exert greater pressure upon schools to remediate academic shortcomings (Smith & Shepard, 1987) than would parents of lower socioeconomic status. These factors would thus produce higher rates of retention in schools of High socioeconomic status. This scenario is not supported by the data obtained in this school district.

A third scenario is that the highest retention rate would be found in schools of Middle socioeconomic status. In these schools one might expect to find the widest range of socioeconomic status and student achievement. In this environment the low achieving student would be conspicuous, teachers may again have concerns regarding the ability of the low achieving students to keep up (Smith, 1989), and teachers may feel the greatest need to achieve a degree of homogeneity within the classroom (Smith, 1989). Although our findings are consistent with this line of reasoning, the data do not allow us to rule out the existence and impact of other factors which may bear on the decision to retain or promote.

A plausible and possibly optimistic explanation for the wide variety of retention practices observed in this school district is that decades of research regarding the ineffectiveness and negative impacts of retention are bearing fruit. The variation in practices could be caused by the uneven diffusion and acceptance of this body of knowledge. It is indeed possible that we are in a period of transition as research findings begin to inform practice.

Hess and Greer (1987) suggest, and anecdotal reports corroborate, that principals are key players in the decision to promote or retain, and that the principals' attitudes regarding retention affect teachers' recommendations. Further analysis of the role of the principal could yield valuable insights, and research into the variation by school socioeconomic status of teacher and principal attitudes toward retention would be appropriate.

This line of reasoning suggests that a fruitful avenue for research lies in the use of ethnographic methods in which an observer follows individual retention cases and analyzes the interactions of the participants. Such research could delineate the source and patterns of influence which affect the decision to retain or promote. Finally, we recommend that comprehensive data regarding retention be collected, just as statistical information is collected regarding special education and students with disabilities.

Our conclusion is that the characteristics of individual students cannot account for the variation in retention rates observed across schools in this district. This conclusion is disturbing; no one would argue that the probability of a student repeating a grade should depend on the school catchment area in which a student resides. The fact that rates vary widely within school SES categories also suggests that factors external to the student impact the probability of retention. It is difficult to construct a scenario in which individual student characteristics could explain why the highest-retaining school of Middle SES retained students at a rate 33 times that of the lowest-retaining Middle SES school. It is our hope that additional research illuminates these issues, and that the patterns of retention we observed are an indication that educational practice is being informed by research findings.

References

Abidin, R. R., Golladay, W. M., & Howerton, A. L. (1971). Elementary school retention: An unjustifiable, discriminatory, and noxious educational policy. *Journal of School Psychology, 9*, 410-417.

Bossing, L., & Brien, P. (1980). *A review of the elementary school promotion-retention dilemma.* Murray, Kentucky: Murray State University. (ERIC Document Reproduction Service No. ED 212-362).

Byrnes, D. (1989). Attitudes of students, parents, and educators toward repeating a grade. In L. A. Shepard & M. L. Smith (eds.), *Flunking grades: The politics and effects of retention* (pp. 16-33). Philadelphia: Falmer Press.

Byrnes, D., & Yamamoto, K. (1986). Views on grade repetition. *Journal of Research and Development in Education. 20*(1), 14-20.

Gallup, A. M. (1986). The 18th annual Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 68*, 43-59.

Gallup, A. M. (1983). The 15th annual Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 65*, 33-47.

Gastright, J. F. (1989, March). *The nation reacts: A survey of promotion/retention rates in 40 urban school districts.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA (ERIC Document Reproduction Service No. ED 307-714).

Hersh, L. R. (1988). *Correlates of elementary school retention: A case study.* Horseheads Central School District, Horseheads, NY. (ERIC Document Reproduction Service No. ED 294-675).

Hess, G. A., & Greer, J. L. (1987). *Bending the twig: The elementary years and dropout rates in the Chicago Public Schools.* Chicago, IL: Chicago Panel on Public School Policy and Finance. (ERIC Document Reproduction Service No. ED 287-951).

Holmes, C. T. (1989). Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard & M. L. Smith (eds.), *Flunking Grades: The politics and effects of retention* (pp. 16-33). Philadelphia: Falmer Press.

Jackson, G.B. (1975): The research evidence on the effects of grade retention. *Review of Educational Research, 45*(4), 613-635.

Mantzicopoulos, P., Morrison, D. C., Hinshaw, S. P., & Carte, E. T. (1989). Nonpromotion in kindergarten: The role of cognitive, perceptual, visual-motor, behavioral, achievement, socioeconomic, and demographic characteristics. *American Educational Research Journal, 26*, 107-121.

Morris, D. R. (1991, April). *Structural patterns and change in grade retention rates: An aggregate analysis of data from a large urban school district, 1982-1989.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 334-337).

Rose, J. S., Medway, F. J., Cantrell, V. L., & Marus, S. H. (1983). A fresh look at the retention-promotion controversy. *Journal of School Psychology, 21*, 201-211.

Safer, D. J. (1986). Nonpromotion correlates and outcomes at different grade levels. *Journal of Learning Disabilities, 19*, 500-503.

Safer, D., Heaton, R., & Allen, R. P. (1977). Socioeconomic factors influencing the rate of non-promotion in elementary schools. *Peabody Journal of Education, 54*, 275-281.

Shepard, L. A., & Smith, M. L. (1989). Academic and emotional effects of kindergarten retention in one school district. In L. A. Shepard & M. L. Smith (eds.), *Flunking Grades: The politics and effects of retention* (pp. 79-107). Philadelphia: Falmer Press.

Smith, M. L. (1989). Teachers' beliefs about retention. In L. A. Shepard & M. L. Smith (eds.), *Flunking Grades: The politics and effects of retention* (pp. 79-107). Philadelphia: Falmer Press.

Smith, M. L., & Shepard, L. A. (1987). What doesn't work: Explaining policies of retention in early grades. *Phi Delta Kappan, 69*(2), 129-134.

Tomchin, E. M., & Impara, J. C. (1992). Unraveling teachers' beliefs about grade retention. *American Educational Research Journal, 29*, 199-223.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 23-30

Using Research Results on Class Size To Improve Pupil Achievement Outcomes

C. M. Achilles
Eastern Michigan University

Patrick Harman
SERVE

Paula Egelson
SERVE

Beginning in 1991-1992 a local school system spent local funds to implement class-size reductions (grades 1-3) to 1:15 as a way to improve the early schooling of primary pupils. Evaluations at grade 3 (1994) showed statistically significant differences ($p \leq .01$) between **classes** *of 1:15 (n=17) and of 1:25 (n=26) in matched "experimental" (n=4) and "control" (n=7) schools with effect sizes ranging from .65 to .73. Based on these results, education leaders fully implemented 1:15 in grades 1-2 with local funding in all elementary schools for the 1994-1995 and 1995-1996 school years.*

Introduction

Pupil achievement remains a major issue in today's schools. Since funds for schooling are not unlimited, school leaders must choose how they deploy available resources. One criterion for fund use might be, "Do the things we choose to do have a *sound* basis in research?" This paper reviews briefly the results of one local school district's commitment to use research results related to class size (1:15) to improve schooling for its elementary pupils.

Since the early research on class size and student achievement (starting in the 1920s) and the more recent debates on class size such as the Glass and Smith (1978) meta-analysis and responses to it by Robinson (at the Education Research Service, or ERS, 1978 and 1980) in the mid-1980s, statewide studies of class size and student learning have been added to the equation. Two of these that have published results are Project Prime Time in Indiana (Mueller, Chase, & Walden, 1988) and Project STAR (Student Teacher Achievement Ratio) in Tennessee (Achilles, Nye, Zaharias, & Fulton, 1993; Word et al., 1990). Writing for the American Academy of Arts and Sciences (AAAS), Mosteller (1995) noted STAR's size (over 8,000 pupils), experimental design (random assignment of pupils and teachers to one of three conditions), and longitudinal nature (grades K-3), and critiqued STAR in gracious terms:

> Project STAR, a study of the educational effects of class size in the state of Tennessee, is one of the great experiments in education in United States history. Its importance derives in part from being a statewide study and in part from its size. But more important yet is the care taken in the design and execution of the experiment. (p.1., Executive Summary)

In addition, STAR has led to "subsidiary," "ancillary," and "related" studies derived from and/or building on the STAR initiative. (Table 1 has a sample of these types of studies.) The composite results from studies such as these are providing a solid base for education decisions in at least 11 states (Bracey, 1995).

Youngsters entering school today are different from youngsters entering public schools only a few years ago in terms of poverty, linguistic differences, single-parent homes (or homeless), abuse, and other problems that impede traditional education processes. (Flaxman & Passow, 1995; Hamburg, 1992; Hodgkinson, 1991, 1992; Mitchell & Cunningham, 1990): Society changes, but often schools -- or those who work in schools -- do not change to accommodate the changing clientele. Some educators simply refuse to use what research has shown

---

Charles M. Achilles is a Professor of Educational Leadership at Eastern Michigan University. Patrick Harman is an evaluation specialist with SERVE in Greensboro, NC. Paula Egelson is a program specialist with SERVE. Correspondence regarding the paper should be addressed to Dr. C. M. Achilles, Educational Leadership, Eastern Michigan University, Ypsilanti, MI 48197.

to work (e.g., Glickman, 1991). Some school leaders, however, give cause for optimism as they study and then use the results of education research to improve student outcomes. Project Challenge in Tennessee provides one example of using class-size research for broad-scale improvement for youth in poverty (Achilles, Nye, Zaharias & Fulton, 1995). Rural Burke County, North Carolina (NC) has also achieved student gains by applying the results of class-size research, but not without considerable local expense. Why did Burke County educators elect to use class-size research, and what were some of the results?

Table 1
Samples of Studies Derived From and Building Upon the STAR Initiative Classed as "Subsidiary" (directly from STAR), "Ancillary" (building on and using STAR database) and "Related" (triggered by STAR results and usually involving STAR researchers).

CATEGORY, TITLE & PURPOSE *	*DATE(S)*	*AUTHOR(S) OR PUBLICATION*
Subsidiary Studies		
• Lasting Benefits Study to follow STAR pupils	1989-Present	Nye et al., 1994
• Project Challenge (TN)	1989-Present	Nye et al., 1994
• Participation on Grades 4, 8	1990, 1994	Finn, 1989 Finn and Cox, 1992
Ancillary Studies (Use or extend STAR data. Some of these are dissertations.)		
• Retention in Grade	1994	Harvey, 1994
• Achievement Gap	1994	Bingham, 1993
• Value of K in Classes of Varying Sizes (test scores)	1985-1989	Nye et al., 1994-1995
• School-Size and Class-Size Issues	1985-1989	Nye, K., 1995
• Random v. Non-Random Pupil Assignment and Achievement	1985-1989	Zaharias, 1995
• Class Size and Discipline in Grades 3,5,7	1988,1990,1992	In Process, Hibbs.
• Outstanding Teacher Analysis (top 10% of STAR teachers)	1985-1989	Bain et al., 1992
Related Studies		
• Success Starts Small: Grade 1 in Chapter 1 (1:14, 1:23) Schools	1993-1995	Achilles et al., 1994

* This list is not complete. It provides samples of the types of studies done. Not all authors appear in the references in the exact way listed here. This table appears in several STAR reports in substantially this form.

Prior Research: A Brief Synopsis of STAR's Class-Size Findings

> This research leaves no doubt that small classes have an advantage over larger classes in reading and mathematics in the early primary grades. (Finn & Achilles, 1990, p. 573)

Project STAR, and STAR's subsidiary, ancillary, and related studies confirm a positive class-size effect on pupil learning. Tennessee's Lasting Benefits Study (LBS) has shown that pupils originally in small (1:15) classes in STAR have retained their achievement advantage over pupils in regular (1:25) classes at least through grade 7 (the last year that data are available and analyzed), although the achievement differences between the groups are less pronounced in grade 7 than in grade 3, the end of the STAR treatment (Nye, Boyd-Zaharias, Fulton, Achilles, & Pate-Bain, 1994). There are, as Finn and Achilles (1990) noted, other questions that need to be studied and answered, but the basic question, "Do smaller classes positively affect student achievement in early primary grades?" seems clearly answered in the affirmative. Some local districts are now reducing class sizes in elementary grades.

Context and Setting

Scenic Burke County is the 21st largest of North Carolina's 121 school systems with approximately 12,800 students in two high schools, four middle schools, and 14 elementary schools. Interstate 40 cuts through the center of the 511 square-mile county that contains considerable tax-exempt properties, such as State parks, Pisgah National Forest, and State facilities. Growth in some industries (e.g., poultry) has brought the mixed blessings of increased economic growth and an influx of workers whose children may need extra educational help.

Older education facilities have recently been remodeled or renovated and are well maintained. The condition of facilities seems to belie the fact that in the 1993-1994 school year Burke County received over $200,000 in low-wealth reimbursement -- funds awarded a North Carolina school system that taxes itself highly but still cannot generate required funds for the education program. (In 1993-1994 there were 1500 Chapter 1-eligible youngsters, about 12% of the total system membership, but funds to serve only 825.) The system also exceeds its "cap" on handicapped students and there are funds to serve only about one-third of the academically-gifted youngsters. Yet, this system houses some education opportunities of great merit as Burke County residents have gone "the extra mile," to support exemplary programs at local expense.

The Class-Size Initiative

In 1991, Burke County educators recognized a need to help early primary pupils experience success in schools. A study group comprised of teachers, administrators, parents and community members analyzed the problem and reviewed possible research-based solutions, including available results from Project Prime Time (Indiana) and Project STAR (Tennessee) showing the benefits of low teacher-pupil ratios in early primary grades. The Tennessee and Indiana studies supported prior research and a major class-size meta-analysis of class size and student achievement (Glass & Smith, 1978).

Reports from the Education Research Service (ERS, 1978; 1980), a research review of class size and student achievement (Robinson, 1990), policy statements and analyses (Tomlinson, 1988, 1990) and commentaries on class size and student achievement (e.g., Slavin, Madden, Karweit, Livermon, & Dolan, 1990) have presented differing viewpoints, data, and interpretations of class-size research. Although the debate continues, it is less that class-size reduction increases student achievement -- the question that STAR answered -- than that other approaches may help more. Indeed, STAR researchers said that not only did reduced class size *cause* increased achievement but that small class sizes could *facilitate* improved instruction -- that class size was not only a causal variable, but also a facilitative variable (Word et al., 1990).

After reviewing the class-size research results, the Burke County study group recommended that school leaders initiate a pilot study of reduced class size (1:15) generally following the Tennessee and Indiana models. If preliminary assessments showed the value of the 1:15 treatment in grade 1, the study would proceed, providing an opportunity to study not only year-to-year pupil progress, but also the effects of the full treatment by using results of the End of Grade (EOG) tests given statewide to grade 3 pupils. Initial findings from grades 1 and 2 using local indicators indicated that reduced-class students outperformed regular-class students in reading and math (Burke County Schools, 1993). Thus, the study continued through grade 3 (1993-94) with pupils in four elementary schools in classes of about 1:15 and pupils in 10 other schools in classes of about 1:25. Table 2 shows the progression of the study, including that after the pilot-study group exited a grade, all schools at that grade level (grades 1 and 2) then adopted the 1:15 framework based on the early year-to-year "matched pairs" analyses.

Table 2
Burke County, North Carolina, 1:15 Study Showing Progression (1991-92 to 1993-94) of Students in Small (S or 1:15) and Regular (R or 1:25) Class Conditions Distributed Throughout the 14 Elementary Schools

		Year of the Study with Schools in Condition			
Grade	Type	91-92	92-93	93-94	94-95
1	S	4	14	14	14
	R	10	0	0	0
2	S	0	4	9	14
	R	14	10	5	0
3*	S	0	0	4	4
	R	14	14	10	10

* Grade of assessment with statewide End-of-Grade (EOG) tests. After pupils moved through the pilot test all 14 schools became 1:15 in grades 1 and 2.

Local costs to support the class-size initiative and the subsequent installation of 1:15 in grades 1 and 2 have exceeded 5 million dollars over four years (1991-1992 to 1994-1995). Have Burke County educators, leaders, and citizens made a judicious decision regarding class-size reduction and pupil achievement? Has reduced class size (1:15) aided student achievement on tests? Although the final answer awaits graduation of the first cohort of 1:15 students (2003), there are some preliminary answers.

Burke County educators requested assistance from South Eastern Regional Vision for Education (SERVE), the regional educational laboratory for the Southeast. Because SERVE's mission is to promote and support the improvement of educational opportunities for all learners in the Southeast, staff seek opportunities to collaborate with reform projects. The Burke County initiative represented such an opportunity, and SERVE personnel assisted with the evaluation (1994).

Design

Burke County educators began the 1:15 effort with a pilot study in four schools in 1991-92. The remaining 10 elementary schools contained larger classes (about 1:25). (Table 2 describes the grade progression of 1:15 classes from 1991-1992 through 1993-1994). For the grades 1 and 2 analyses, local evaluators matched individual pupils selected from the 1:15 classes with pupils from the 1:25 classes based on sex, socio-economic status (SES), teacher experience, and race. After the first year, pre-test score information became part of the match. The instruments were the D.C. Heath reading series pre/post tests given about six months apart. Other indicators of success included teacher and parent comments, reviews of portfolios, and pre/post-tests of arithmetic.

Based on prior year results and pre-testing in a current year, pupils were "rematched" each year in the pilot stages. This reduced the ability to detect a difference over more years than one due to the rematching process, but it did show differences in group scores in a particular year. By keeping all 1:15 classes in 4 schools, 10 schools were available to serve as "control" schools for the analysis of differences between groups based upon the End of Grade (EOG) tests given *statewide* at grade 3. This EOG analysis was done first with Spring, 1994 test results.

Rather than pre-determining which schools of the 10 that did not implement 1:15 would be the "control" schools, SERVE evaluators tested all 10 for the best "match" with the 4 schools that implemented the 1:15 treatment. Variables used to determine the match between *schools* with 1:15 and with 1:25 were race, SES as percent of Chapter 1 and percent pupils on free/reduced lunch (FL), percent parents with less than a high school education (<HS), and the school's prior year's (1992-1993) average EOG test scores for reading and math. Results of the comparison between the control schools selected (n=7) and reduced-class schools (n=4) are shown in Table 3. There were no statistically significant differences in the small differences shown in the comparisons.

Table 3
Comparison of Reduced-Class (1:15) and Control (1:25) Schools in Burke County to Establish Comparability
There Were No Significant Differences

		Percent in Category				1993 EOG Scores	
Condition	(n)*	white	<HS	Ch 1	FL	Read	Math
Control	7	88.7	25.9	24.9	36.8	141.9	138.9
Reduced	4	87.6	29.6	16.7	40.8	141.2	139.4

* Four schools with 17 classes and 7 schools with 26 classes. The prior year's EOG scores show comparability *before* 1:15 pupils arrived. (<HS = less than high school education. FL = Free or reduced lunch. Ch.1 = Chapter 1)

There were several reasons for selecting more control schools than reduced-class schools. The first was that the reduced-class schools generally had more than one reasonably comparable school with which to be matched. Consequently, it made sense to include all schools that were "like" the experimental schools. In this way, results of the statistical analyses do not depend on the vagaries of choosing one control school rather than another; selecting seven control schools rather than four should provide a better representation of the counterfactual.

The second reason was an issue of sample size. Since these analyses were on teacher (or class) means and the number of teachers was relatively small, it is beneficial to increase sample size if feasible. In this way, mean differences between the 1:15 schools and 1:25 schools would be easier to detect. To accommodate the fact that pupils in a class are not independent learners (their progress is tied both to the teacher and to the rest of the class), the analyses of differences depended on the *average* scores of 17 classes (1:15) in 4 schools and of 26 classes (1:25) in 7 schools. The EOG test scores (1993-1994) were the criteria for determining the efficacy of 1:15 as a treatment.

Results

Initially, SERVE researchers re-ran the original local evaluation groups using the original matching of students from 1991-1992 and 1992-1993 but using as the new criterion measures the 1993-1994 EOG test results rather than the local benchmarks. This new analysis, referred to as the "matched pairs" analysis, grouped the matched pairs so that the class averages could be used to check if the early analysis and the new analysis processes provided similar results.

Results are provided here (briefly) for matched-pairs analyses at two points: third graders re-matched as first graders and third graders re-matched as second graders. Those summaries are followed by the main focus of this study, the analyses of differences of EOG test results of 17 classes of 1:15 in 4 schools and of 26 classes (1:25) in 7 schools. Rather than showing only single-year comparisons as the matched pairs analyses might, the main analysis shows cumulative effects, grades 1-3, of the 1:15 classes.

The Matched Pairs Process and Analyses

Beginning in 1991-1992 in grade 1, local personnel established a "matched pairs" process by matching pupils in 1:15 with pupils in a non-treatment school. During the pilot test, evaluators used local benchmarks, but when the 1994 grade-3 EOG results for all pupils became available, it was possible to re-establish the "matched pairs" groups and to compare their results post hoc on the EOG tests. That is, using the grade 3 EOG tests it was possible to evaluate math and reading scores of grade 3 pupils who had been re-matched as first graders and those who were re-matched as second graders. All comparisons shown in Table 4 favor the reduced (1:15) classes over the control; three reach significance at $p \leq .05$ and one is not significant (NS) with $p = .13$. Effect sizes (ES) range from .37 to .56. The ES is computed as the mean of the experimental group minus the mean of the control group divided by the pooled standard deviation (SD). Thus, ES's here are interpreted in SD units between classes, *not* between students.

In computing the results in Table 4, the re-matched students were grouped with their teachers to form "classlets," and the weighted averages of these classlets were used as the units for analysis in the tests for differences. The aggregating of pupil scores around a teacher was one way to adjust for the issue of independence of measures.

Table 4
Third Grade Math and Reading Means, Probability Levels, and Effect Sizes for the Two Third-Grade Matched Pairs Statistical Analyses Using Pairs Matched as First Graders and Pairs Rematched as Second Graders

	Treatment	Mean	*p* Level	ES
Math				
Pairs rematched	Reduced	147.10	.134	N/A
as first graders	Control	144.02		
Pairs matched	Reduced	146.38	.041*	.45
as second graders	Control	140.96		
Reading				
Pairs rematched	Reduced	146.36	.022*	.37
as second graders	Control	144.07		
Pairs matched	Reduced	148.00	.012*	.56
as first graders	Control	143.66		

* Significant at $p \leq .05$

The Grade 3 EOG Analyses for 1:15 and 1:25 Classes

Analyses of the test-score differences of the 17 reduced classes (1:15) and 26 control classes (1:25) in the 4 and 7 schools respectively resulted in statistically significant ($p \leq .05$) differences in reading and in math favoring the 1:15 condition. (See Table 5). A variable called "academic" was created as the composite of the 1994 EOG grade 3 tests in reading, math, science and social studies. This variable was standardized to a mean of 50 and SD of 10. This analysis also used class means as the unit of analysis. The 1:15 classes outperformed the 1:25 classes ($p \leq .02$) on the academic variable, too.

Results of the statistical tests clearly and consistently favor the 1:15 treatment in Burke County at the EOG testings in grade 3 for reading, math, and "academic" at $p \leq .05$. Based on these results the Burke County Board of Education voted in July, 1994 to extend the 1:15 treatment to all 14 elementary schools, grades 1 and 2 in the county system.

Table 5
Grade 3 Analyses of 1994 EOG Reading and Math Scores of 17 Reduced Classes (1:15) in 4 Schools with Scores of 26 Control Classes (1:25) in 7 Matched Schools

Small Classes (1:15)	Reading	Math
Cases (n)	17	17
Mean Score	144.46	142.87
Standard deviation	2.92	4.61
Control Classes (1:25)		
Cases (n)	26	26
Mean Score	142.45	139.75
Standard deviation	3.19	4.01
t-Test Information		
Difference (Mean X - Mean Y)	2.02	3.12
t-statistic	2.09	2.38
Degrees of freedom	41	41
Probability of t (Two tailed)	0.04	0.02
ES: Reduced Class (1:15)	0.65	0.73

Discussion

Results of the Burke County 1:15 study are similar to those found in Tennessee (e.g., Word et al., 1990) and in Indiana (e.g., Mueller et al., 1988) and support the findings of these two studies. They also are similar to a small-class study, Success Starts Small, conducted in 1993-1994 (Achilles, Kiser-Kling, Owen, & Aust, 1994). That there is some consistency in the findings of these studies of reduced class size (i.e., the benefits of 1:15 over 1:25) may be attributed, at least in part, to the *length* of these studies, the size of the studies, and the fact that the studies have been implemented in the early primary grades -- the first years of schooling for the pupils in the studies. These research refinements corrected deficiencies in earlier class-size research.

Criticisms of early class-size studies included a) that the treatments were too brief to provide substantive benefits, b) that studies were conducted with pupils in later years of schooling [STAR results suggest strongly that the (1:15) treatment must occur in K or 1 to be of value (Word et al., 1990), and Robinson (1990) concluded that the small-class treatment was most beneficial in K-3], c) that studies were not well controlled, or d) that researchers typically used the pupil as the unit of analysis rather than the class average and that this confounded class size and teacher effects. Some projects designed to aid pupil achievement in *specific* subject areas (such as Reading Recovery or Success for All) build on small classes or use the ultimate small-class treatment of one-to-one tutoring. In the national movement for charter schools, the authors have not yet encountered a discussion of a charter school that builds upon *large* classes. Few home-schooling efforts employ large classes.

Policy applications of experimental research such as Tennessee's Project Challenge help validate educational change based on class-size adjustments (Achilles et al., 1995; Nye, Achilles, Boyd-Zaharias, & Fulton, 1994; Nye, Achilles, Boyd-Zaharias, Fulton, & Wallenhorst, 1994). The Burke County results based on three years of 1:15 treatment extend and support class-size reductions as sound education policy for early grades.

Burke County education leaders are building upon a growing database of support for their actions as some states are moving ahead with class-size initiatives (Bracey, 1995). Indeed, this evaluation of the class-size effort in Burke County shows gains at grade 3 of over .5 ES.

A good early start in school seems important for student achievement and later success in school. Replications of experimental findings in "regular" or non-experimental settings such as Burke County or in Project Challenge should help education leaders base important decisions on the best research available at the time. We should avoid, as Glickman (1991) says, "Pretending not to know what we know."

References

Achilles, C. M., Kiser-Kling, K., Owen, J., & Aust, A. (1994). *Success Starts Small*. Final Report. A Small-Grant School-Based Research Project. Chapel Hill, NC. The Board of Governors of the University of North Carolina System.

Achilles, C. M., Nye, B. A., Zaharias, J. B., & Fulton, D. (1995, April). *Policy use of research results: Tennessee's Project Challenge*. Paper at the American Educational Research Association, San Francisco, CA.

Achilles, C. M., Nye, B. A., Zaharias, J. B., & Fulton, D. (1993). Creating successful schools for all children: A proven step. *Journal of School Leadership, 3*(6), 606-621.

Bain, H. P., Achilles, C. M., Zaharias, J. B., & McKenna, B. (1992). Class size does make a difference. *Phi Delta Kappan, 74*(3), 253-256.

Bingham, S.* (1993). *An examination of small class as a "gap reduction" strategy for achievement differences in groups of students, K-3.* Unpublished Ed.D. dissertation. University of North Carolina, Greensboro.

Bracey, G. (1995) Research oozes into practice: The case of class size. *Phi Delta Kappan, 77*(1), 89-90.

Burke County Schools. (1993). *Reduced class size.* Morganton, NC: Author.

Education Research Service or ERS. (1978). *Class size: A summary of research.* Arlington, VA: Author.

Education Research Service or ERS. (1980). *Class size research: A critique of a recent meta-analysis.* Arlington, VA: Author.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*(3), 557-577.

Finn, J. D., & Cox, D. (1992). Participation and withdrawal among fourth-grade pupils. *American Educational Research Journal, 29*(1), 141-162.

Flaxman, E., & Passow, A. H. (1995). *Changing populations: Changing schools,* (eds). Chicago IL: University of Chicago Press.

Glass, G. V., & Smith, M. L. (1978). *Meta-analysis of research on the relationship of class size and achievement.* San Francisco: Far West Laboratory for Educational Research and Development.

Glickman, C. (1991). Pretending not to know what we know. *Educational Leadership, 48*(8), 4-10.

Hamburg, D. A. (1992). *Today's children.* New York: Time Books, Random House.

Harvey, B.* (1993). *An analysis of grade retention for pupils in K-3.* Unpublished Ed.D. study. University of North Carolina, Greensboro.

Hibbs, B. F.* (In Process). Relationships among discipline factors and early student placement in small (1:15), regular (1:25) and regular-with aide classes.

Hodgkinson, H. (1991). Reform vs reality. *Phi Delta Kappan, 73*(1), 8-16.

Hodgkinson, H. (1992). *A demographic look at tomorrow.* Washington, DC: Institute for Educational Leadership.

Mitchell, B. and Cunningham, L., (Eds.), (1990). *Educational leadership and changing contexts of families, communities, and schools.* Chicago, IL: University of Chicago Press.

Mosteller, F. (1995). *The Tennessee study of class size in the early school grades.* Cambridge, MA: American Academy of Arts and Sciences, Initiatives for Children. (136 Irving Street, 02138).

Mueller, D. J., Chase, C. I., & Walden, J. D. (1988). Effects of reduced class sizes in primary classes. *Educational Leadership, 45,* 48-50.

Nye, B. A., Achilles, C. M., Boyd-Zaharias, J., Fulton, B. D., & Wallenhorst, M. (1994). Small is far better. *Research in the Schools, 1*(1), 9-20.

Nye, B. A., Achilles, C. M., Boyd-Zaharias, J., & Fulton, B. D. (1994). *Project Challenge: Fourth-year Summary Report.* Nashville, TN: Center of Excellence for Research in Basic Skills, Tennessee State University.

Nye, B. A., Achilles, C. M., & Bain, H. P. (1994-95). The test-score "value" of kindergarten for pupils from three class conditions, at grades 1, 2 and 3. *National Forum of Educational Administration and Supervision Journal, 12*(1), 3-15.

Nye, B. A., Boyd-Zaharias, J., Fulton, B. D., Achilles, C. M., & Pate-Bain, H. (1994) *The Lasting Benefits Study: Fourth-year report.* Nashville, TN: Center of Excellence for Research in Basic Skills, Tennessee State University. (Also reports in 1993, 1992, and 1991).

Nye, K. (1995). *The effect of school size and the interaction of school size and class type on selected student achievement measured in Tennessee elementary schools.* Unpublished Ed.D. Dissertation, University of Tennessee, Knoxville.

Robinson, G. E. (1990). Synthesis of research on the effects of class size. *Educational Leadership, 47*(7), 80-90.

Slavin, R. E., Madden, N. A., Karweit, N. J., Livermon, B. J., & Dolan, L. (1990). Success for all: First-year outcomes of a comprehensive plan for reforming urban education. *American Educational Research Journal, 27*(2), 255-278.

Tomlinson, T. M. (1988). *Class size and public policy: Politics and panaceas.* Washington, DC: US Department of Education, Office of Educational Research and Improvement.

---

*These studies are extending the STAR, LBS and Challenge data as a way to try to understand the impact of either a small class or a full-time teacher aide on the conditions being studied. They are a part of a series of dissertations.

Tomlinson, T. M. (1990). Class size and public policy: The plot thickens. *Contemporary Education, LXII*(1), 17-23.

Word, E., Johnston, J., Bain, H., Fulton, B., Zaharias, J., Lintz, N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/teacher achievement ratio (STAR): Tennessee's K-3 class size study.* Final report and final report summary. Nashville, TN: Tennessee State Department of Education.

Zaharias, J. B.* (1993, December). *The effects of random class assignment on elementary students' reading and mathematics achievement.* Unpublished Ed.D. dissertation. Tennessee State University, Nashville.

Author Notes

The authors wish to acknowledge the Burke County (NC) Board of Education, prior superintendent Mr. Carlos Hicks, Dr. Richard Jones, Dr. Wayne Honeycutt, Dr. Guy McBride, Mr. Tony Stewart and other Burke County administrators for their dedication to this project. Further, the researchers acknowledge the Burke County teachers for their help. The regional educational laboratory (South Eastern Regional Vision for Education or SERVE) provided time, resources, personnel and expertise to assist in the study. The evaluation effort continues. Nothing herein is to be construed as an official position of SERVE or of the US Department of Education which provides major funding for SERVE.

The authors thank three anonymous reviewers for comments helpful in the revision process. Any errors of omission or commission, however, are the authors' own.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 31-38

Biology Students' Beliefs about Evolutionary Theory and Religion

Anne Sinclair and Beatrice Baldwin
Southeastern Louisiana University

High school students are bringing increased religious bias to the biology classroom due to the resurgence of religious conservatism which teaches there must be a forced-choice between "creationist" Biblical literalism and evolutionary theory. Students' beliefs were shown to interfere with their ability to objectively view scientific evidence, especially when they involve deeply ingrained religious teachings which are counter to the information being presented. Those who held less conservative views more readily accepted the credibility of scientific evidence supporting the theory of evolution and were better able to reconcile their religious beliefs with the scientific concepts. All of the students who rejected the credibility of evolutionary theory gave as their reason opposing religious views. A number of misconceptions that students held about evolutionary theory were identified.

The controversy continues between evolutionary biologists and creationists, those who believe in a literal interpretation of the Bible and hold that the earth is relatively young and that all living things were created in a short span of time. Their beliefs are contradictory to evolutionary theory which posits that the earth is billions of years old and living species have evolved over long periods of time (Tatina, 1989). McInerney (1991) contends that creationists argue from an inverted epistemological hierarchy which equates a scientific theory to something akin to a guess, rather than perceiving it as having a compelling conceptual framework that explains a multitude of evidences.

The creationists are displaying greater determination and political sophistication since the Supreme Court declared unconstitutional a Louisiana law (*Edwards vs. Aquillard*, 1987) requiring equal time for the teaching of creationism and evolution. The law in question was struck down because its formulation was viewed as unconstitutional, not because the teaching of creationism was deemed unlawful. Conservative religious groups interpreted this ruling as leaving the door open for the teaching of creationism, and across the United States they have been bringing considerable pressure on local school boards to include creationism in the curriculum, even to the exclusion of evolutionary theory (Zimmerman, 1991).

Many who make school decisions do not understand the theory of evolution and are "easy prey from purveyors of pseudoscientific claptrap" (McInerney, 1991, p. 5). Zimmerman's survey of school board presidents in Ohio confirms this, for only 1.8% were able to select the correct description of evolutionary theory and 53% of those responding favored the teaching of creationism in public schools. This pattern appears to be emerging in the United States. In a recent poll of the general public, 74%-86% indicated that creationism should be included in the curriculum (Zimmerman, 1991).

Because of the resurgence of religious conservatism, students often bring emotional and religious biases to biology class (Kaplan, 1994). They often believe there is a forced "either/or" choice between their religious faith and evolutionary theory. Undue pressure is brought to bear on these students because of the seemingly forced dichotomy which requires that one reject the theory of evolution and accept a literal interpretation of the Biblical account of creation. Certainly not all religious faiths insist upon this dichotomous choice. Scott and Cole (1985) observed, "Christian religions, not based upon Biblical literalism, long ago made peace with evolution" (p. 28).

Biology students' tolerance for ambiguity is low, especially among those who think at the intuitive and concrete levels (Lawson & Worsnop, 1992; Nelson, 1986; Scharmann, 1990). It is difficult for students to accept that scientific theories are tentative and that choices must be made as to which theory offers the best scientific explanation. Perhaps the most threatening ambiguity experienced by biology students is that even though evolutionary theory withstands rigorous scrutiny and presently offers the best explanation for the diversity of life, there are areas of controversy among respected evolutionary biologists. It is important for students to realize that these disagreements are not about the basic

---

Anne Sinclair is a former high school biology teacher who is presently an Assistant Professor of Science Education at Southeastern Louisiana University. Beatrice Baldwin is the Assistant to the Vice President for Research, Planning, and Development at Southeastern Louisiana University. Please address correspondence to Anne Sinclair, Department of Teacher Education, Southeastern Louisiana University, SLU Box 936, Hammond, LA 70402 or e-mail at Asinclair@selu.edu.

conceptual framework of evolutionary theory, but concern specific processes of change. "There are no alternatives to evolution as history that can withstand critical examination. Yet we are constantly learning new and important facts about evolutionary mechanisms" (Dobzhansky, 1973, p. 129).

It is not likely that students' misconceptions will be significantly changed during a high school instructional unit on evolutionary theory, but it is crucial that the concepts be presented in such a way as to promote rational thinking and insistence on supporting evidence (Lawson & Worsnop, 1992). Despite their religious convictions, students must be made aware that science offers the most objective approach for answering questions about the physical universe (Scott & Cole, 1985).

The purposes of this study were twofold. The first was to determine the relationship between students' religious beliefs and their ability to objectively view the scientific evidence supporting evolutionary theory. It was hypothesized that those who held less conservative religious beliefs would more likely accept the credibility of the theory's explanations. This hypothesis is consistent with the research findings of Lawson and Worsnop (1992). The second purpose was to identify specific misconceptions that students held about the theory of evolution after a thorough coverage of the supporting evidences.

Method

Data were collected from 212 high school biology students in 11 classes in two high schools, each with a population of approximately 1000. Both schools, located within the same school district in southeast Louisiana, have well-equipped science laboratories. One hundred and twelve females (52.8%) and 100 males (47.2%) participated in the study. The demographic characteristics of the sample reflect the community at large which has a range of socioeconomic and cultural diversity. One hundred and fifty two of the students were white (71.7%), 59 were black (27.8%), and one was Hispanic (<.5%). The mean age of the students was 16 years and 5 months. Based upon the most recent census records, approximately 35% of the families in the area are active in conservative religious organizations.

A three-week unit on evolutionary theory was taught during the spring of 1994. The curriculum guides for high school biology, both at the state and parish levels in Louisiana, require a comprehensive coverage of the major tenets of evolutionary theory. The key concepts covered during the instructional unit were the origin of life on earth, Darwin's theory of evolution and the mechanisms of natural selection, the mechanisms of speciation, modern evidences of evolution, and primate evolution. The unit on evolutionary theory was preceded by a study of classification systems and followed by biological reproduction.

The students were instructed using diverse methodologies including large group presentations of content which involved questioning and discussion, small group projects and discussions, laboratory investigations and experiments using the inquiry approach, and viewing of videos and slides of paleontological, morphological, biochemical and genetical evidences which support the theory. The instructional unit was culminated with a field trip to a natural science museum which displays an excellent collection of fossils and preserved specimens.

To insure that all students in the study received similar instruction, the two teachers targeted the same instructional objectives in their lesson plans, utilized the same textbook (Towle, 1989) and laboratory activities, and incorporated the same outcome measures in their assessment of the students' mastery of the learning objectives.

At the beginning of the instructional unit, the teachers explained to the students that they were evolutionary biologists and they were also theists. In both instructors' classrooms the climate was open and accepting to the sharing of opinions and beliefs. Students were apprised numerous times that their ideas would be heard without censure. On several occasions creationist beliefs were brought up by the students. The teachers allowed the students to share their ideas, but reminded them that when any claims were made, because this was a science classroom, scientific evidence must be presented to document their beliefs.

Upon the completion of the three week unit on evolutionary theory, the students were asked to respond anonymously to a questionnaire which targeted their understanding of the theory of evolution. They were also requested to describe their religious beliefs in relation to the theory. Twelve of the questions were multiple choice and four of the items asked the students to explain or describe their beliefs by responding in narrative. A pre-measure was not administered because of the controversial nature of the issues being considered. Asking students to state their beliefs about evolutionary theory and creationism prior to the instructional unit could have threatened the internal validity of the measure by calling their attention to the controversial nature of the topics, perhaps promoting close-mindedness to the scientific evidences which would be presented during instruction.

Results

Descriptive statistics were calculated for all of the multiple choice items and the results are reported in Tables 1-4. Qualitative data taken from the narrative responses to questions on the survey were summarized and are reported in the sections which follow.

Open-mindedness and Change in Thinking

Nearly all (97.2%) of the students indicated that they were either very open or somewhat open to new ideas (question 1 - see Table 1), and 80.7% responded that they had changed their thinking about evolution during the unit (question 3). Students were then asked to explain their responses. Cited below are representative comments that were made by those who said their thinking had changed during the instructional unit on evolutionary theory:

> "I feel more educated now."
>
> "I now believe the theory is very possible."
>
> "The theory is very believable."
>
> "It opened my mind to new ideas and possibilities, but also confused me."
>
> "Before, I thought evolution did not occur at all because I felt it taught we came from apes. Now I realize that there are other things that don't go against my beliefs such as change in species over a long period of time."
>
> "I believed about 50% of it."
>
> "I can't prove it wrong so I might as well believe it."
>
> "It makes me stronger in believing the Bible instead of the evolutionists."

Of the 41 students who said their thinking had not changed, 100% stated that their conservative religious beliefs would not allow them to consider evolution as a credible theory.

When describing how their opinions could be influenced (question 2), two of the options were selected most often. Approximately 37% of the students were persuaded primarily by rational ideas supported by evidence, while nearly 35% said they were influenced the most by a person who could be trusted to tell the truth about the new ideas. Parents' and friends' beliefs were the least chosen options (27.8%).

Table 1
Questionnaire Items Pertaining to Open-mindedness and Change in Thinking

Question	Response
1. How open minded to new ideas do you consider yourself?	
a. very open-minded	82 (38.7%)
b. somewhat open-minded	124 (58.5%)
c. generally, my opinions are firm and unchanging	6 (2.8%)
2. When presented with new ideas, which of the following would influence you the most?	
a. The person who is sharing the new ideas is trusted to tell the truth as they know it.	74 (34.9%)
b. The new ideas are rationale and/or have sound scientific support.	79 (37.3%)
c. Friends' beliefs about the new ideas.	42 (19.8%)
d. Parents' beliefs about the new ideas.	17 (8.0%)
3. Has studying this unit on the scientific theory of evolution changed your thinking or opinions? Please explain your response.	
a. Yes	171 (80.7%)
b. No	41 (19.3%)

Understanding of Scientific Theory and the Theory of Evolution

Descriptive statistics are reported in Table 2 for questions which pertain to their understanding of evolutionary theory. Nearly 62% of the students selected the correct definition for a scientific theory (question 4) and 81.1% were able to choose the appropriate description of the theory of evolution (question 5). It is noteworthy that even though such a large percentage of the students selected the correct response, only 23.6% correctly identified the primary mechanisms that result in change (question 6), and close to the same number (24.1%) were able to indicate the correct description of Darwin's (1962) theory of "Survival of the Fittest" (question 7). The remaining 75.9% selected Lamarckian explanations. Clearly, the students did not have an understanding of the mechanisms of speciation even though they had spent three weeks studying the supporting evidences.

Table 2
Questionnaire Items Pertaining to Understanding of Scientific Theories and the Theory of Evolution

Question	Response
4. Which of the following is the best definition of a scientific theory?	
a. A belief that something is true based upon a large body of evidence, yet more evidence is being sought.	131 (61.8%)
b. A law or principle that has been proven to be true.	25 (11.8%)
c. An educated guess that something is true.	43 (20.3%)
d. An idea which has little evidence or support.	13 (6.1%)
5. Which of the following is the most scientific description of the theory of evolution?	
a. Life on earth is constant and unchanging.	13 (6.2%)
b. Life on earth has changed in the past but is now constant and unchanging.	10 (4.7%)
c. Life on earth has changed, is presently changing and is predicted to continue changing in the future.	172 (81.1%)
d. Life on earth has changed in the past but is not likely to change in the future because of human intervention and control.	17 (8.0%)
6. Which of the following would best explain what causes species to change?	
a. Organs and structures which are not needed are lost.	53 (25.0%)
b. Changes in the earth's climate.	44 (20.8%)
c. DNA mutations which allow organisms to compete more successfully.	50 (23.6%)
d. Some organisms run out of food and die while others survive because they migrate to new territories.	65 (30.6%)
7. Which of the following best describes Darwin's theory of *Survival of the Fittest*?	
a. Organisms adapt in order to survive.	128 (60.4%)
b. Dinosaurs died because they ran out of food due to drastic changes in the climate.	16 (7.5%)
c. Giraffes have long necks so that they can reach the leaves on tall trees.	17 (8.0%)
d. The most competitive organisms survive long enough to reproduce fertile offspring.	51 (24.1%)

Strongest and Weakest Evidences of Evolutionary Theory

Nearly 40% of the students believed that fossil records (question 8 - see Table 3) offered the strongest supporting evidence for evolution. The next highest selection was variation in organisms with 30.7% choosing this option. A smaller number of students (29.7%) chose the more complex evidences such as biochemical, genetical or embryological similarities, even though these topics were given equal coverage during instruction.

Table 3
Questionnaire Items Relating to Strongest and Weakest Evidences Supporting Evolutionary Theory

Question	Response
8. What do you feel is the strongest evidence scientists give to support the theory of evolution?	
a. Fossils of species that are now extinct.	84 (39.6%)
b. Variations in organisms that are alive today (*example*: races of humans).	65 (30.7%)
c. Homologous structures (*example*: bird's wing and human's arm).	10 (4.7%)
d. Biochemical similarities between species.	14 (6.6%)
e. Embryological similarities between species.	14 (6.6%)
f. Genetical similarities between species.	25 (11.8%)
g. Other (please describe)__________	0 (0.0%)
9. What do you feel is the weakest argument scientists give to support the theory of evolution?	
a. Life originated from a type of simple cell.	85 (40.1%)
b. Fossils of extinct organisms.	24 (11.3%)
c. The Geological/Biological Time Table describing changes of the earth and species over time.	25 (11.8%)
d. Species which are alive today have common ancestors.	18 (8.5%)
e. Complex organisms such as humans evolved from simpler species.	49 (23.1%)
f. Other (please describe) __________	11 (5.2%)
10. Does anything concern or trouble you about the scientific theory of evolution? Please explain your response.	
a. Yes	121 (57.1%)
b. No	91 (42.9%)

Most of their selections concerning the weakest evidences scientists give to support the theory of evolution (question 9) centered around whether life originated from single cells (40.1%) and whether humans evolved from other species (23.1%). When asked whether anything troubled them about evolutionary theory (question 10), 57.1% described specific concerns. Twenty-seven of the students (12.7%) who responded in the affirmative expressed confusion about human ancestry in relation to other primates. One comment seemed to epitomize their remarks: "I am still confused about man's relationship to the monkey -- did or did not man evolve from the monkey or from something else?" Other comments were:

> "The whole thing scares me."
>
> "It (evolutionary theory) is really too complicated."
>
> "If man came from monkeys, why are there still monkeys?"
>
> "I don't believe I will ever understand evolution to its full extent, but I don't believe I want to either."
>
> "What keeps us from going backwards? "

Religion and Evolutionary Theory

Only 6.1% of the students felt that religious teachings and evolutionary theory were not in disagreement (question 11 - see Table 4), yet 68.9% said that one could believe in evolution and God (question 12). One student stated, "I thought scientists couldn't believe in evolution and God but now I know they can. Evolution is more than how we got here, it's how we've changed." Another wrote, "At first I didn't believe that evolution and God could go together. Now I think maybe God might have caused evolution. No one really knows, but it's fun to think about." Other representative comments made regarding religious beliefs were:

> "I can't believe God made us and believe we came from monkeys -- it goes against the Bible. "
>
> "I despise that anyone would try to mislead someone else away from God. It has made me realize how lost evolutionists are. I will never accept evolution!"
>
> "It (evolutionary theory) goes against what I've been taught to believe."
>
> "It's so confusing for a Christian to accept."
>
> "How do we know there really is a God?"
>
> "Evolution proves God and the Bible wrong."

When asked to describe their present feelings and beliefs about the scientific theory of evolution (question 13), the comments were varied, ranging from vehemently opposed, to neutral, to mildly accepting, to endorsing. A sampling of the students' responses are cited below:

> "People who believe in this theory are heading down the wrong path."
>
> "It's a lie."
>
> "Not sure, don't know, and really don't care."
>
> "Still don't know what to believe, but I do believe God put us here."
>
> "I need more evidence -- it could be true."
>
> "I now understand that everything evolved from something else."

Table 4
Questionnaire Items Pertaining to Religion and Evolutionary Theory

Question	Response
11. Which of the following best describes your opinion about how the scientific theory of evolution and religious teachings are related?	
a. I do not feel they are in disagreement.	13 (6.1%)
b. I feel they are somewhat in disagreement	67 (31.6%)
c. I feel that they disagree.	79 (37.3%)
d. I am not sure whether they agree or disagree.	53 (25.0%)
12. Do you feel that a person can accept the theory of evolution and also believe God? Please explain your response.	
a. Yes	146 (68.9%)
b. No	66 (31.1%)
13. Please describe your present feelings and beliefs about the theory of evolution.	
14. What are some things your biology teacher could have done to assist you in better understanding the scientific theory of evolution?	

Chi-square analysis (see Table 5) indicated a significant difference in the ability to reconcile scientific and theistic beliefs between students who could correctly describe evolutionary theory and students who could not ($\chi^2_1 = 7.20$, $p = .007$). Approximately seventy-three percent of the students who chose the correct description of evolutionary theory felt that one could accept this theory and still believe in God and the Bible, while only 51.3% of those who could not correctly describe the theory felt that these beliefs were reconcilable.

Table 5
Summary of Chi-Square Analysis Comparing Correct Response to Question 5 and Attitudes toward Reconciling Beliefs (Question 12) (N= 212)

	Do you feel that a person can accept the theory of evolution and also believe in God? (Question 12)		
Which of the following is the most scientific description of the theory of evolution? (Question 5)			
	Yes	No	Total
Chose Correct Response	126	46	172
Chose Incorrect Response	20	20	40
Total	146	66	212

* $\chi^2_1 = 7.20$, $p = .007$

Recommendations for Their Biology Teachers

The students were asked to identify ways their biology teacher could have assisted them in better understanding the theory of evolution (question 14). These responses also represented a wide range of opinions. Seventeen students (8%) expressed feelings relating to creationism. One said, "I don't care what the teacher or anyone else says, I know what the Bible says." Other comments were:

> "(Teacher's name) should have included a lesson on Creationism so we could have had a basis for comparison."

> "Creationism supposedly forces religion on people, but the belief in evolution is also a type of religion. Both ideas should be presented equally."

> "My concern is that evolution is being taught in my school and not Creationism. Creationism is not just a belief; there are many scientific facts to back it up."

> "I need a lot more explanation. This was all new to me and it frightened me."

> "Go slower and give me time to absorb all this."

Some wanted more labs (6%) and others felt the unit needed to last longer so that they could do some additional investigation (7%). Eighteen percent wanted more details about how species had evolved, especially relating to the geological time table. Many of the comments (58%) indicated that their teachers had done a good job in explaining the theory of evolution and they felt no instructional changes were needed. One student expressed that the teacher had been extremely supportive and "did not put anybody down because of their opinions." Another said, "I don't know how (teacher's name) had enough patience to put up with us."

Conclusions and Implications

From the results of this study, as well as others (Lawson & Worsnop, 1992; Scharmann, 1990), we can conclude that students' beliefs interfere with their ability to objectively view scientific evidence, especially when they involve deeply ingrained religious teachings which are counter to the information being presented. Our hypothesis that those who held less conservative religious beliefs would more likely accept the credibility of evolutionary theory was strongly supported. All of the students who rejected the credibility of evolutionary theory gave as their reason opposing conservative religious views.

Another objective of this study was to identify misconceptions held about evolutionary theory. It was obvious that many did not have an understanding of the complexities of the theory. Even after three weeks of instruction, a large number of the students did not comprehend the mechanisms of speciation and natural selection. They also selected the more observable and less complicated evidences which support evolutionary theory. It would appear that students often find the global tenets of the theory difficult to comprehend.

Biases may have prevented some of the students from even considering scientific evidence. They seemed to have been inordinately preoccupied with the speciation of *Homo sapiens*. Even though the theories of primate evolution were clearly addressed citing specific paleontological, biochemical and genetical evidences, the students tenaciously held to misconceptions and biases regarding ancestral relationships between humans and primates.

An important concern life science educators must address is how students can be assisted in overcoming beliefs which prevent them from objectively viewing scientific evidence. Scharmann (1993) believes that students need to discuss their beliefs about creation origins and evolution and give reasons why each of the arguments is compelling. This enhances classroom participation and during the process students often come to realize that they are not required to make an "either/or" choice between scientific theories and religion. Students in Scharmann's 1990 study expressed feelings of relief when they found that they were not alone in their confusion.

> Providing the student with a place to stand between two extremes is critical, especially if we are to promote an understanding of the nature of science and why scientists feel that the theory of evolution is such a major cornerstone of the biological sciences. Failure to provide such a position to students leaves them far too often with an uncomfortable choice . . . . (p. 98)

Kemp (1988) contends that biology instructors should intentionally hold up the ideas purported by creationists for scrutiny, for the essence of science is to question and probe, insisting upon validating evidence. If this is done it should be with sensitivity, for deeply felt belief systems are being challenged. Perhaps a better approach would be to address creationists' claims if they are brought up by students. At this point it would be appropriate to investigate whether scientific evidence supports or refutes such claims. Lawson and Worsnop (1992) would agree, for they recommend that biology lessons be planned to promote hypothetico-deductive reasoning, requiring the student to compare alternative hypotheses in much the same way as scientists do.

It would seem that the most compelling approach would be to address students' concerns in an open forum. When this is done, respect and consideration for the students and their deeply felt beliefs are very important. The idea is neither to indoctrinate nor condescend, but to promote an understanding of the uniqueness of scientific inquiry and its insistence on valid and supportable evidence.

Another recommendation is to include evolutionary theory throughout the school year when its relevance can be explained and addressed in context, rather than teach it as a single expanded unit near the end of the academic year. By doing this, students will have more time to process meaning and to make cognitive adjustments when their biases are preventing them from being open-minded. The present curriculum for many biology classes does not allow students enough time nor sufficient varied experiences to process the complex and voluminous information being presented. For many, it is simply overwhelming both academically and emotionally.

This study was descriptive in nature, yet the findings can serve as an impetus for additional research. By identifying specific misunderstandings and biases, instructional strategies can be recommended which assist students as they confront whether they should believe the teachings of the creationists or whether they should accept the evidences presented by evolutionary theory. Science and religion both seek to answer human questions, but from entirely different epistemological bases (Scott & Cole, 1985). Deciding which realm is best suited to answer questions about the origin and development of life on earth presents an enigma to many students. "For most it is a lonely and personal thing" (Stokes, 1989, p. 24).

References

Darwin, C. (1962). *The origin of the species* (6th ed., originally printed 1872). New York: Macmillan.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher, 35*(3), 125-129.

Kaplan, G. R. (1994). Shotgun wedding: Notes on public education's encounter with the new Christian Right. *Phi Delta Kappan, 75*, 1-12.

Kemp, K. W. (1988). Discussing creation science. *The American Biology Teacher, 50*(2), 76-81.

Lawson, A. E., & Worsnop, W. A. (1992). Learning about evolution and rejecting a belief in special creation: Effects of reflective reasoning skill, prior knowledge, prior belief and religious commitment. *Journal of Research in Science Teaching, 29*, 143-166.

McInerney, J. D. (1991). A biologist in wonderland: The Texas biology textbook adoption hearings. *The American Biology Teacher, 53*(1), 4-5.

Nelson, C. E. (1986). Creation, evolution or both? A multiple model approach. In Robert W. Hanson (Ed.), *Science and creation: Geological, theological, and educational perspectives*. New York: Macmillan Publishing Company.

Scharmann, L. C. (1990). Enhancing and understanding of the premises of evolutionary theory: The influence of a diversified instructional strategy. *School Science and Mathematics, 90*, 91-100.

Scharmann, L. C. (1993). Teaching evolution: Designing successful instruction. *The American Biology Teacher, 55*(8), 481-485.

Scott, E. C., & Cole, H. P. (1985). The elusive scientific basis of creation "science." *The Quarterly Review of Biology, 60*(1), 21-30.

Stokes, W. L. (1989). Creationism and the dinosaur boom. *Journal of Geological Education, 37*, 24-26.

Tatina, R. (1989). South Dakota high school biology teachers and the teaching of evolution and creationism. *The American Biology Teacher, 51*(5), 275-280.

Towle, A. (1989). *Modern biology*. New York: Holt Rinehart & Winston.

Zimmerman, M. (1991). The evolution-creation controversy: Opinions of Ohio school board presidents. *Science Education, 75*, 201-214.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 39-45

Principal Leadership Style, Personality Type, and School Climate

Dawn T. Hardin
Northeast Louisiana University at Monroe

The purpose of this study was to determine the relationships among three factors: principal leadership style, personality type, and school climate; and additionally to determine if the relationships differed according to urban or rural designation and/or school configuration. No relationship was found between the personality type and leadership style of principals in this investigation. Furthermore, neither principal personality types, leadership styles, nor teacher rated climate dimensions discriminated between rural and urban school geographical designations. In studies where such differences were found, many of the rural and urban school samples utilized also represented vast differences in student population size. This study, which utilized systems with student populations similar in size, questions the strength of characteristics associated with rural/urban designations.

The quality of leadership provides the foundation for the success and endurance of an organization (Barnard, 1938; Conger, 1992; Daniels, 1994; Fiedler, 1967). If educators endeavor to create schools that are generative and proactive places, then attention should be focused upon the leadership behavior of principals (Blank, 1987; Boyer, 1983; Richard & Basom, 1993; Sizer, 1984). Effective principal leadership encourages and supports the professional behavior of the teaching staff. Ultimately, this contributes to the success of the learning environment in terms of climate and academic achievement (Hurley, 1994; McCleary, 1983; Spade, Vanfossen, & Jones, 1985). Considering the vital link between leadership and achievement, it becomes essential for educators to focus on school leadership improvement. As noted by Strange (1993) and Glasman (1994), school leadership improvement is an intricate, complicated endeavor because effective principal behavior is not easily determined due to its various roles, dimensions, and responsibilities.

Research has attempted to identify and quantify the qualities and/or behaviors that comprise effective principalship (Keedy, 1992; Lueder, 1983; Pitner & Charter, 1984; Vornberg, 1992). These attempts have contributed to clusters of knowledge that, although fragmented, offer insight into the complex realm of principal effectiveness.

One such cluster, communication, has emerged as a vital component of principal effectiveness. Faculty tend to share ideas among each other and with the administration. Early studies conducted by Wellische, MacQueen, Charrier, and Duck (1978) and more recently by Keedy (1992) found that where these faculty communication patterns are both frequent and successful, there exists an increased likelihood of higher student achievement, enhanced teacher satisfaction and greater principal effectiveness.

In conjunction with communication, another cluster of knowledge which appears to impact school climate concerns the principal's style of leadership or management. Shreeve and Others (1984) early noted that principals who foster participatory management systems within schools tend to increase the job satisfaction of their teachers. More recently, Vornberg (1992) propounded that excessive control demonstrated by principals was undesirable in school settings. Vornberg further emphasized the need for positive principal and teacher communication by his assertion that methods for principal selection include an assessment of a candidate's attributes and skills in the areas of interpersonal relations and group dynamics.

Further scrutiny of effective principal leadership research reveals that other clusters of knowledge have evolved from studies conducted by Blake and Mouton (1981), Warrick, (1981), Yukl (1981), and others. These and other researchers note relationships between personality types, leadership styles, and behavior. In addition, Lueder (1983) proposed a link between psychological type and principal behavior by his conclusion that the manner by which a principal responds to problem situations can be characterized by his or her

---

Dawn T. Hardin is an assistant professor in the Department of Educational Leadership and Counseling at Northeast Louisiana University at Monroe. Correspondence regarding this paper should be addressed to Dr. Dawn T. Harding, 306 Strauss Hall, Dept. of Educ. Leadership and Counseling, Northeast Louisiana University, Monroe, LA 71209-0230.

BEST COPY AVAILABLE

individual psychological type. Lueder's assertion has been supported by others who maintain that principals, as well as other leaders, vary by their identified leadership styles, perceived leader behaviors, and personality types (Dobbs, 1989; Guthrie & Reed, 1986; Kroeger & Thuesen, 1988).

According to Purkey and Rutter (1987), the environmental contexts of schools will continue to emerge as an area for additional research. More recently, Witte and Walsh (1990) found that social and organizational environments of suburban, urban, and rural schools d affect teaching staff perceptions and principal leadership. Therefore, in studies of school effectiveness and climate, it is suggested that environmental contexts be considered (Hannaway &Talbert, 1993).

Furthermore, school configuration appears to affect effectiveness and climate. In concordance with the much debated leadership substitute theory, some researchers assert that principal leadership differs among dissimilar school configurations. It has been proposed by many that high school principals are not "lone leadership figures" due to the school configuration's divided tasks and departmentalization, whereas elementary principals typically exert sole leadership and authority because they serve as the only administrative figure at the school site (Gallagher, Riley, & Murphy, 1986; Kerr, 1977; Kerr & Jermier, 1978).

While knowledge concerning the numerous aspects of principal effectiveness abounds, it is difficult to ascertain whether principal effectiveness is absolute or if it serves as a function of the environmental context and/or school configuration in which it resides. Therefore, further research is needed to understand the nature and functions of principal effectiveness and school climate more completely.

Statement of the Problem

The research purpose of this investigation was to determine the relationships between the principal's leadership style and personality type, and school climate. Furthermore, this study sought to determine if the relationships differed according to the school's urban or rural geographical designation and/or the school's configuration.

Methodology

Subjects

This investigation utilized responses from principals and teachers from one urban school district and two rural parish districts. The principal sample consisted of the 34 principals who responded out of the three-district population of 50 principals. The teacher sample which represented 20% of the three-district population of 1,383 teachers was established using systematic sampling of alphabetized faculty roles provided by each participating principal. The 34 principal responses and 166 teacher responses resulted in a return rate of respectively 68% and 12% of the principal and teacher populations in the three selected school districts. In addition, the response rate was proportional to the size of the district and number of schools in the area.

Principals representing three distinct school configurations participated in the study. Twenty-two of the principal respondents were principals of elementary schools; eight participants were from secondary schools; four were from K-12 schools. Two additional respondents were excluded from the analysis due to missing data. Participating urban principals ranged in age from 33.62 to 59.08 with a mean age of 50.94 while the rural principals ranged in age from 33.43 to 57.85 with a mean age of 46.84. All principals did not specify gender. Of the gender responses reported, 3 urban principals were female; 6 were male. According to rural responses, 8 principals were female and 12 were male.

Instrumentation

This investigation utilized three instruments. Principals recorded their responses to the *Leadership Behavior Analysis II* [LBAII] and the *Myers-Briggs Type Indicator* [MBTI] while the teachers' perceptions of leadership and climate were assessed by the *School Assessment Survey* [SAS].

The *Leadership Behavior Analysis II* is a self-scoring assessment which resulted in the principal's analysis of his or her own behavior producing an identification of one of the following primary leadership styles:

S1 - High Directive, Low Supportive Behavior
S2 - High Directive, High Supportive Behavior
S3 - High Supportive, Low Directive Behavior
S4 - Low Supportive, Low Directive Behavior.

Nye (as cited in Zigarmi, Edeburn, & Blanchard, 1991) in his test/retest reliability study of the *Leadership Behavior Analysis II* reported an Index of Stability coefficient of .72. Other studies conducted for the purpose of establishing construct and predictive valid-ity used comparisons between the *Leadership Behavior Analysis II Other* and the *Multi-Level Management Survey*. A significant relationship ($p <. 0001$) was found in all comparisons except for Subscale 4, Expertise, which was significant at .0004 (Zigarmi, Edeburn, & Blanchard, 1991). For the purposes of this study, it was believed that the *Leadership Behavior Analysis II* was a reasonable measurement of a principal's primary leadership style.

BEST COPY AVAILABLE

The *Myers-Briggs Type Indicator* was designed according to Jung's theory of type. The instrument contains items which sort people into groups. The instrument was used to classify the principals according to the following four separate dichotomies or indices:

Extraversion - Introversion
Sensing - Intuition
Thinking - Feeling
Judging - Perceiving.

The *Myers-Briggs Type Indicator* measures personality dimensions where the polarities of each are viewed as strengths. Extensive reviews by Carlyn (1977) and Carskadon (1979) of intercorrelation, reliability, and validity studies indicate that the *Myers-Briggs Type Indicator* is related to variables such as personality measures, and other scales such as selected Strong Vocational Blank scales and the Edwards Personal Preference Schedule. In conclusion, the instrument can be considered an adequately reliable self-report inventory and was considered a reasonable determination of principal personality type for this study.

The *School Assessment Survey* acquired information concerning school climate, resulting in profiles indicating each individual school's and groups of schools' scores indicating climate for the following nine dimensions:

1. Goal Consensus (GCON),
2. Facilitative Leadership (PLEAD),
3. Centralization of Influence - Classroom Instruction (CINT),
4. Centralization of Influence - Curriculum and Resources (CINF),
5. Vertical Communication (VCMN),
6. Horizontal Communication (HCMN),
7. Staff Conflict (HCFT),
8. Student Discipline (DISC), and
9. Teaching Behavior (TEACH).

Alpha results for the dimensions of the *School Assessment Survey* range from alpha coefficients of .76 to .96 denoting its use as a reliable measurement of school climate for this investigation. Data for the *School Assessment Survey* are summarized across teachers for each of the nine dimensions. For all but one dimension, the central tendency of teacher responses within a given school is calculated. To complete the climate analysis, the dispersion of responses for the dimension of goal consensus is depicted in the summary score (Wilson, Firestone, & Herriott, 1985).

Procedures

In each school district, principals completed the *Leadership Behavior Analysis II* during a regularly scheduled principals' meeting. The principals were then given a *Myers-Briggs Type Indicator*, a return envelope, and instructions to complete the instrument and return it within five days. Each principal supplied the researchers with teacher rolls which were randomized to provide a 20% teacher sample. Mailed to this teacher sample were packets containing a cover letter, a *School Assessment Survey*, and a return envelope. Teachers had 10 days to respond. Once evaluations of the research instruments were completed, the data were analyzed on site using discriminant function analysis.

Results

Data initially were analyzed using discriminant function analysis with *School Assessment Survey* scores entered to predict school type, rural or urban. The Wilks' lambda associated with the extracted variate indicated that these variables lacked sufficient discriminating power to justify interpretation, Wilks' lambda = .93, X^2 (2) = .31, $p > .05$.

In an effort to examine the link between *Myers-Briggs Type Indicator* personality types and *Leadership Behavior Analysis II* leadership types, Spearman's rank-order correlation coefficients were calculated between each *Myers-Briggs Type Indicator* scale and *Leadership Behavior Analysis II* type. Because only one respondent was classified as S4, the coefficients were calculated excluding the single aberrant individual. Results were not significant.

Data then were subjected to a second stepwise discriminant function analysis with eight teacher rating scores from the *School Assessment Survey* entered to predict school configuration type, either elementary, secondary, or K-12. This procedure evaluates the contribution of each variable by assessing the increase in Rao's V for each variable entered into the equation. In this sense, an optimal set of predictor variables is achieved and presumably reflects the most relevant factors which separate the groups.

Two discriminant functions were extracted from the data (see Table 1). Chi-square calculated on Wilks' lambda suggested only the first function was statistically significant, Wilks' lambda = .42, X^2 (10) = 25.35, $p < .01$. With the first function removed, the data retained insufficient discriminating power to justify the interpretation of the second function, Wilks' lambda = .82, X^2 (4) = 5.87, $p > .05$. Interpretation of the canonical correlations associated with the variates also suggests the first function is of greater substantive importance, Rc = .70. This function accounts for the large proportion of the variance, 81%.

Table 1
Results of Discriminant Analysis Predicting School Type

Discriminant Functions Derived

Function	Canonical Correlation	Wilks' Lambda	*df*	Chi Squared
1	.70	.42	10	25.35**
2	.43	.82	4	5.87

Discriminating Variables

Step	Variable Entered	Variable Removed	Rao's V	Wilks' Lambda
1	HCMN		14.33	.68 **
2	CINF		22.78	.58 **
3	HCFT		26.39	.52 **
4	VCMN		30.75	.48 **
5	CINT		34.23	.43 **
6	TEACH		39.56	.40 **
7		HCMN	36.63	.42 **

* $p < .05$ ** $p < .01$.

Five of the eight predictor scores -- Vertical Communication, Staff Conflict, Centralization of Influence - Classroom Instruction, Centralization of Influence - Curriculum and Resources, and Teaching Behavior -- were entered into the final discriminant equation. Horizontal Communication, having entered at step 1, was removed at step 7. This indicates the contribution made by this variable, although important, adds little to group separation after other variables were entered into the discriminant function.

The extracted functions then were rotated to simple structure using varimax criteria. Rotated correlations between the discriminating variables and the canonical discriminant functions are presented in Table 2. Interpretation of the correlations suggested the first function tapped a bipolar dimension measuring communication, both horizontal and vertical, and centralization of influence on curriculum and resources. The rotated correlations between Horizontal Communication and Vertical Communication and the variate were positive and above the .40 criterion. Centralization of Influence - Curriculum and Resources generated a negative correlation with function 1 after rotation. The remaining variables were associated with function 2. Only Staff Conflict, Centralization of Influence - Classroom Instruction, and Student Discipline, however, had rotated correlations that met the interpretation criterion. Because this variate appears primarily artifactual, the associations with this function are equivocal.

Table 2
Varimax Rotated Correlations Between Discriminating Variables and Canonical Discriminant Function

Variable	Function 1	Function 2
VCMN	.59	.13
CINF	-.55	-.02
HCMN	.44	.03
HCFT	.07	-.71
CINT	.21	.62
DISC	.11	.56
Group Centroids		
Elementary	.67	.10
Secondary	-.95	-.69
Combined	-1.77	.85

Note. Only function 1 was statistically significant

Location of the group centroids (see Table 2) suggests maximal group separation was achieved among the three groups along this first and most potent function. The elementary teachers were primarily associated with the positive end of the dimension. The centroids of the remaining groups, secondary and K-12, were located toward the negative end of the dimension with the K-12 group generating the largest negative centroid. Separation along the second dimension was less satisfactory. The unimpressive separation of centroids along this dimension along with the summary statistics associated with it indicated that it not be interpreted.

Univariate *F*-ratios and the lambda associated with each of the variables in the predictor set are presented in Table 3. As indicated, only three variables were statistically significant. As might be expected these variables all loaded with the first extracted variate. Inspection of the cell means also presented in Table 3 reveals that the highest ratings on the two communication scales, vertical and horizontal, are observed among the elementary school teachers with the lowest ratings obtained from the K-12 teachers. For the Centralization of Influence - Curriculum and Resources scale, highest means were observed among the K-12 teachers with lowest ratings obtained from the elementary teachers. In all cases the secondary teachers generated means between the other groups.

Table 3
Univariate Results of Discriminating Variables

	Wilks' Lambda	F	Elementary M(SD)	Secondary M(SD)	Combined M(SD)
HCMN	.68	7.16**	2.62(.42)	1.98(.65)	1.80(.78)
VCMN	.74	5.32*	1.68(.58)	1.12(.43)	.97(.26)
HCFT	.90	1.78	.91(.49)	1.13(.60)	.56(.28)
CINT	.88	2.07	-.87(.33)	-1.16(.41)	-.89(.34)
CINF	.77	4.54*	1.44(.48)	1.87(.39)	2.06(.45)
DISC	.94	.99	3.25(.49)	3.00(.64)	3.41(.43)
TEACH	.99	.22	64.38(11.54)	60.86(16.83)	64.15(11.47)
PLEAD	.99	.02	3.79(1.06)	3.74(1.12)	3.86(.76)

* $p < .05$. ** $p < .01$.

Results of the classification phase of the analysis are presented in Table 4. As shown in Table 4, overall classification accuracy was impressive, 74%. This percentage necessarily is inflated because neither a jackknife nor partial cross-validation was performed. The limited sample size in this study, however, prohibited the applications of these more conservative methods. The classification accuracy, nonetheless, represented a substantial increase over the prior probability of .33 and suggested the solution may be adequately separating the groups.

Table 4
Results of Discriminant Classification Procedures

Actual Group Membership	n	Predicted Group Membership 1	2	3
1 Elementary	22	82%	5%	14%
2 Secondary	8	38%	50%	13%
3 Combined	4	25%	0%	75%

Note. The overall classification accuracy is 74%.

In order to determine whether the present sample generated a particular *Myers-Briggs Type Indicator* profile, X^2 was calculated on each *Myers-Briggs Type Indicator* pair. Interestingly, principals were rather evenly split between introversion and extraversion. On the remaining scales, however, a significant pattern was detected. Of the total sample, more principals rated themselves as Sensing rather than Intuitive $\{X^2 = (n = 33) = 25.00, p < .05\}$, Thinking rather than Feeling $\{X^2 = (n = 27) = 9.00, p < .05\}$, and Judging rather than Perceiving $\{X^2 = (n = 31) = 18.78, p < .05\}$. This apparent homogeneity of personality type on three of the four personality factors might explain the lack of relationship between *Leadership Behavior Analysis II* and *Myers-Briggs Type Indicator*.

Conclusions

No relationship was found between the personality type and leadership style of principals in this investigation. Furthermore, neither principal personality types, leadership styles, nor teacher rated climate dimensions discriminated between rural and urban school geographical designations. In literature where such differences were found, many of the rural and urban school samples utilized also represented vast differences in student population size (Blank, 1987; Hannaway & Talbert, 1993). This study, which utilized systems where student populations were similar in size (a population mean of 405.25 for rural schools and a population mean of 558.55 for urban schools), questions the strength of rural/urban designations.

Do rural and urban schools exist in different educational worlds created by unique organizational and administrative factors ascribed by an environmental context, or are these differing realms created by the powerful impact of size? This question is of utmost importance considering the present trends of rural consolidation and urban decentralization. Recently, numerous rural communities have conducted consolidation studies to investigate the possibility of providing in a cost effective manner additional course offerings, extracurricular activities, and enhanced learning opportunities for rural students. Many of these communities struggle to offer small numbers of students a curricula which meets state standards, community expectations, and college and university entrance requirements. Consolidation appears to present a viable solution to this problem for small community schools with duplicating programs.

In contrast to the trend of rural consolidation, urban schools appear to move in an opposite direction using decentralizing efforts such as neighborhood schools and site-based management to create student-centered responsive educational environments. As small rural schools become larger and large urban schools become smaller, additional studies are needed to determine the degree of discrimination and interaction between school size and environmental context. Further research especially incorporating "small" urban and "large" rural schools would add considerable insight as well as lead to a more specific clarification of urban and rural school differences (Hardin, Cage, & Santana, 1995).

Data did suggest that the communication patterns in the three school configurations differed. On the vertical and horizontal communication scales elementary schools reported the highest ratings whereas K-12 schools reported the lowest. Yet K-12 schools reported the highest rating for the centralization of influence in terms of curriculum and resources. Interestingly, secondary school ratings consistently fell between the other school groups. Although these findings contribute to this small cluster of knowledge regarding school configuration, studies involving larger samples are needed to clarify the distinct characteristics of each configuration type, and accordingly to contribute to a better understanding of school configuration types. Once identified, the negative and positive characteristics of each configuration type could be addressed to enhance the leadership and climate in all schools.

References

Barnard, C. I. (1938). *The functions of the executive.* Cambridge, MA: Harvard University Press.

Blake, R. R., & Mouton, J. S. (1981). The exercise of effective leadership. *Journal of Experiential and Simulation, 3*(1), 3-16.

Blank, R. K. (1987). The role of principal as leader: Analysis of variation in leadership of urban high schools. *Journal of Educational Research, 81*(2), 69-80.

Boyer, E. (1983). *High school: A report on secondary education in America.* New York: Harper and Row.

Carlyn, M. (1977). An assessment of the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 41*(5), 461-471.

Carskadon, T. G. (1979). Behavioral differences between extraverts and introverts as measured by the Myers-Briggs Type Indicator: An experimental demonstration. *Research in Psychological Type, 2,* 78-82.

Conger, J. A. (1992). *Learning to lead: The art of transforming managers into leaders.* San Francisco, CA: Jossey-Bass Publishers.

Daniels, A. C. (1994). *Bringing out the best in people: How to apply the astonishing power of positive reinforcement.* New York: McGraw Hill, Inc.

Dobbs, R. L. (1989). The relationship between leadership effectiveness and personality type for a group of urban elementary school principals (Doctoral dissertation, Memphis State University, 1988). *Dissertation Abstracts International, 49,* 3564-A.

Fiedler, F. E. (1967). *A theory of leadership effectiveness.* New York: NcGraw-Hill.

Gallagher, K. S., Riley, M., & Murphy, P. (1986). Instructional leadership in the urban high school -- Whose responsibility is it? *NASSP Bulletin, 70*(488), 26-30.

Glasman, N. S. (1994). *Making better decisions about school problems: How administrators use evaluation to find solutions.* Thousand Oaks, CA: Corwin Press.

Guthrie, J. W., & Reed, R. J. (1986). *Educational administration and policy: Effective leadership for American education.* Englewood cliffs, NJ: Prentice-Hall, Inc.

Hannaway, J., & Talbert, J. E. (1993). Bringing context into effective schools research: Urban-suburban differences. *Educational Administration Quarterly, 29*(2), 164-186.

Hardin, D. T., Cage, B. N., & Santana, R. T. (1995, January). *SPACE: The school principal and climate evaluation.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Hurley, J. C. (1994). Become a principal? You must be kidding! *International Journal of Education Reform, 3*(2), 165-173.

Keedy, J. L. (1992, March). *Translating a school improvement agenda into practice: A social interaction perspective to the principalship.* Paper presented at the Eastern Educational Research Association. (ERIC Document Reproduction Service No. ED 348 766)

Kerr, S. (1977). Substitutes for leadership: Some implications for organizational design. *Organization and Administrative Sciences, 8,* 135-146.

Kerr, S., & Jermier, J. (1978). Substitutes for leadership: Their meaning and measurement. *Organization and Human Performance, 22,* 375-403.

Kroeger, O., & Thuesen, J. M. (1988). *Type talk.* New York: Dell Publishing.

Leadership Behavior Analysis II (1991). Escondido, CA: Blanchard Training and Development Inc.

Lueder, D. C. (1983, November). *A study of the relationship between elementary school principals' psychological type and perceived problem-solving strategies.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association. (ERIC Document Reproduction Service No. ED 238 171)

McCleary, L. E. (1983). The urban principal: A new basis for leadership. *NASSP Bulletin, 67*(166), 8-11.

Myers-Briggs Type Indicator (1984). Palo Alto, CA: Consulting Psychologists Press.

Pitner, N. J., & Charter, W. W., Jr. (1984). *Principal influence on teacher behavior: Substitutes for leadership.* Eugene, OR: Oregon University, National Institute of Education. (ERIC Document Reproduction Service No. ED 251 941)

Purkey, S. C., & Rutter, R. A. (1987). High school teaching: Teacher practices and beliefs in urban and suburban public schools. *Educational Policy, 1*, 375-395.

Richard, B. B., & Basom, M. P. (1993). Not "Rambo," not "hero": The principal as designer, teacher and steward. *Educational Considerations, 20*(2), 26-28.

School Assessment Survey (1984). Philadelphia: Research for Better Schools.

Shreeve, W., & Others. (1984). *Job Satisfaction: A responsibility of Leadership.* (Report No. 028 188). Eastern Washington University. (ERIC Document Reproduction Service No. ED 275 638)

Sizer, T. (1984). *Horace's compromise: The dilemma of the American high school.* New York: Houghton Mifflin Company.

Spade, J. Z., Vanfossen, B. E., & Jones, J. D. (1985, April). *Effective schools: Characteristics of schools which predict mathematics and science performance.* Paper presented at Annual Meeting of American Educational Research Association.

Strange, J. H. (1993). Defining the principalship: Instructional leader of middle manager. *NASSP Bulletin, 77*(553), 1-7.

Vornberg, J. A. (1992, April). *Leadership competencies and perceived training effects: Meadows principal improvement program.* Paper presented at the Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 348 773)

Warrick, D. D. (1981). Leadership styles and their consequences. *Journal of Experiential Learning and Simulation, 3*, 155-172.

Wellishe, J. B., MacQueen, A. H., Carrier, R. W., & Duck, G. A. (1978). School management and organization in successful schools. *Sociology of Education, 51*, 211-226.

Wilson, B. L., Firestone, W. A., & Herriott, R. E. (1985). *School Assessment Survey: A Technical Manual.* Philadelphia: Research for Better Schools.

Witte, J. F., & Walsh, D. J. (1990). A systematic test of the effective schools model. *Educational Evaluation and Policy Analysis, 12*, 188-212.

Yukl, G. A. (1981). *Leadership in organizations.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Zigarmi, D., Edeburn, C., and Blanchard, K. (1991). *Research on the LBAII: A Validity and Reliability Study.* Escondido, CA: Blanchard Training and Development Inc.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 47-49

Preservice Teachers and Standardized Test Administration: Their Behavioral Predictions Regarding Cheating

Karyn Wellhousen
The University of New Orleans

Nancy K. Martin
The University of Texas at San Antonio

Sixty-three teacher education students were asked to respond to the idea of cheating as teachers when administering a standardized test to their students. The subjects were asked to clarify and explain their responses. Over half of the subjects said they would cheat under certain circumstances or in specific ways. Among the reasons given were "if it would benefit the students" and "if the test was inappropriate." Ways of cheating considered acceptable included giving hints, rewording test items or directions, and teaching to the exact test given.

As concern regarding the quality of public education has grown, so has the importance of standardized test scores. In spite of other means of evaluation available (i.e., portfolios or performance assessment) the focus on standardized tests remains strong, as does its ripple effect (Canner, 1992; Richards, 1989). Test results have important implications and consequences for schools, as well as the community, the state and the nation. They are used to judge the quality of education and to further political and educational agendas (Canner, 1992). As a result of the tremendous emphasis on test results, school personnel are under pressure to raise test scores (Canner, 1992; Kher-Durlabhji & Lacina-Gifford, 1992; Richards, 1989). Richards (1989) explains, "If [teachers] want to please the principal (who wants to please the superintendent, who wants to please the school board), they have to show that their students can perform on tests . . ." (p. 66).

The emphasis on test results is epitomized by the highly publicized case of a South Carolina teacher who was caught and confessed to providing students with answers to standardized test questions. This highly respected, experienced educator was fired, arrested, and prosecuted. Her certification revoked, she will never teach again (Canner, 1992). It is even more unfortunate that this case does not represent an isolated incident. Evidence suggests school personnel are cheating and the practice is occurring throughout the United States at all levels within the educational system (Cannell, 1989; Canner, 1992; Frisbie & Andrews, 1990; Kher-Durlabhji & Lacina-Gifford, 1992; Ligon, 1985; Perlman, 1985; Richards, 1989; Smith, 1991).

Cheating takes a variety of forms. Cannell (1989) concluded that high test scores are often the result of ". . . lax test security, nonstandard testing practices, deceptive statistics, and misleading impressions, not improved achievement" (p. 5). Teachers have been directed by principals to change students' incorrect answers, hand out copies of the exact test to students well in advance of test administration, write the correct answers on the chalkboard during testing, and remove low-scoring students' answer sheets before they leave the school campus (Canner, 1992; Richards, 1989). Simultaneously, principals are feeling the same pressure from superintendents (Richards, 1989).

Cheating on standardized tests has long-range detrimental effects on us all. Unethical testing practices undermine the public's faith in the educational system and circumvent the schools' accountability to the public at large (Canner, 1992; Popham, 1990). The students suffer most because teachers are modeling unethical behavior. Furthermore, they are distorting information used to make important educational decisions that impact students' lives (Kubiszyn & Borich, 1990; Popham, 1990). With distorted results, we have no way of evaluating what students have and have not achieved or determining their educational needs. As Popham (1990) explains, "Whether it's a crosscut saw, a crescent wrench,

---

Karyn Wellhousen is an Assistant Professor of Early Childhood Education at the University of New Orleans, and she teaches courses in tests and measurements. Nancy K. Martin is an Assistant Professor of Educational Psychology at the University of Texas at San Antonio and also teaches courses in tests and measurements. Please address correspondence regarding this article to Dr. Karyn Wellhousen, Department of Curriculum and Instruction, Lakefront, University of New Orleans, New Orleans, LA 70148.

or an achievement test, a tool can be used properly or improperly" (p. 1). Because sound decisions cannot be based on distorted information, ethical testing practices are of utmost importance.

The literature to date addresses current teachers and their practices when administering standardized tests. Little is known about preservice teachers' expectations and opinions on this subject. Yet, preservice teachers do not exist in a vacuum. They read the professional literature and newspaper reports regarding the emphasis on testing. Therefore, it is likely that they anticipate these pressures. This gives rise to an important question. What predictions do preservice teachers make regarding their future behavior vis-à-vis the administration of standardized tests in their classrooms? Due to the nature of this question, the study was approached in a descriptive, qualitative manner.

Subjects

Subjects in this study were university students in an elementary/early childhood studies program who were enrolled in a required Tests and Measurements course. The age range of students was 19 to 54 years with an average age of 27 years. Sixty-two of the 63 subjects were females. Fifty-nine percent of the subjects were Caucasian, 36% Hispanic, and 5% African-American. The majority of subjects (55%) were Juniors. Fifty percent of the subjects reported having a negative experience of some kind with a standardized test.

Methodology

In the tests and measurement course, standardized testing was a major topic addressed over a period of several weeks. Students were required to review, select, and administer standardized tests. They were familiar with standardization procedures and therefore, understood that deviating from standardized test procedures distorts and invalidates test results. Although the topic of cheating was addressed and discussed in class, the act of cheating was never endorsed or dissuaded by the instructor. The possibility of on-the-job pressure to cheat was addressed.

Students were asked to respond to the following hypothetical questions: "Would you cheat when administering a standardized test? If yes, under what conditions would you cheat? If no, why not?" The word "cheating" was not specifically defined; students self-defined the term. The questions were answered anonymously; therefore, students were not likely to respond to please the instructor.

Responses were placed in one of two categories: "Yes, I would cheat" or "No, I would not cheat." Next, self-explanatory subcategories were established. Subcategories described "how" or "why" subjects would or would not cheat. Responses were independently categorized by the authors. Discrepancies were identified, discussed, and re-categorized accordingly. In some cases, subjects' responses fit into more than one subcategory. (See Table 1.)

Table 1
Summary of Results

	% RESPONDING
YES, I WOULD CHEAT . . . TOTAL N = 39 (62%)	
(WHY?)	
to benefit the children	36
if the test or test items were inappropriate	20
if pressured by others	6
to avoid consequences or receive benefits	6
(How?)	
giving clues/hints	24
rewording test directions, test questions	6
teaching the test	1
no reason given	1
NO, I WOULD NOT CHEAT . . . TOTAL N = 24 (38%)	
(WHY?)	
wouldn't get accurate results about students	16
cheating is morally wrong	8
fear of consequences	5
bad example for students	1
(How?)	
but would give clues	6
but would give practice tests	4
but be careful in describing test results to parents	1
but would teach content covered	1
no reason given	5

Results and Implications

When subjects were asked to predict whether they would cheat when administering a standardized test, 62% of the preservice teachers responded affirmatively. Thirty-six percent stated they would cheat in order to benefit the child. This refers to how the test results will be used. For example, if the preservice teacher knows that a certain test result is required in order for a student to receive a needed service or to be admitted to a special program, they may cheat to "benefit the child." Twenty percent reported that they would cheat if the test or test items were inappropriate. This refers to preservice teachers' interpretation of whether or not a test or specific test items are fair or suitable for the specific group or individual child taking the test. Other reasons given for situations in which cheating was acceptable and the percent of students responding in each category are: if the preservice teacher felt pressured by others (6%), to avoid consequences or receive benefits (6%). Methods describing how they might cheat and the percent responding include: giving clues/hints (24%), rewording test directions or questions (6%), and teaching the test (1%).

Among the subjects who predicted they would not cheat (38%), the following reasons were given as reasons why it would not be appropriate: wouldn't get accurate results about students' knowledge/ability (16%), cheating is morally wrong (8%), fear of consequences (5%), setting bad example for students (1%). Twelve percent of the subjects said they would not cheat when administering a standardized test, but described ways in which they may help students to perform better on tests such as giving clues, giving practice tests, and teaching test content. It is interesting to note that 24% of subjects predicted they would cheat by giving clues, while 6% of the subjects predicting they wouldn't cheat also stated they may give clues.

These results present cause for concern. Over half (62%) of the subjects overtly stated they felt justified in cheating when administering standardized tests to future students. Even though the majority of affirmative respondents qualified their intent to cheat in an effort to benefit their students (36%), 6% of "yes" respondents stated they would cheat if pressured by others, to avoid consequences, or to receive benefits. Equally surprising was the minimal percentage of subjects who stated they would not cheat because it is wrong (8%) or because it would set a bad example for students (1%).

Caution should be exercised when interpreting these results. Self-concept theory suggests a strong relationship between beliefs and actions as individuals strive for consistency between these two (Epstein, 1973). However, because the question requires the self-prediction of future behavior, only time will tell if these personal predictions are borne out. Further, given the limited number of respondents, it is not possible to make broad generalizations from these findings. Finally, the fact that 62 of the 63 participants were female is another limitation of the study. Although the primarily female subject pool is typical of both early childhood and elementary levels, it is not representative of educators at all levels.

In spite of these limitations, it is important to consider the implications of these findings. These results demonstrate the need for further investigation into the opinions of preservice teachers regarding this important ethical issue. Future studies should consider the opinions and predictions of those preservice teachers not enrolled in a tests and measurements course. Subjects' general attitudes toward and personal experiences with standardized tests merit further investigation as these may influence how they view the testing process. The majority of these participants (55%) were classified as college juniors. As students progress through a teacher preparation program, do their opinions evolve and, if so, in what way? In addition to descriptive analyses comparing various levels of preservice educators (juniors to seniors to student teachers), experimental research should follow suit to determine the impact of specific interventions to alter unethical positions.

References

Cannell, J. (1989). *How public educators cheat on standardized achievement tests: The "Lake Wobegon" Report.* Friends for Education, Albuquerque, NM.

Canner, J. (1992). Regaining the public trust: A review of school testing programs. *NASSP Bulletin, 76*, 6-15.

Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist, 28*, 404-416.

Frisbie, D. A., & Andrews, K. (1990). Kindergarten pupil and teacher behavior during standardized achievement testing. *The Elementary School Journal, 90*, 435-448.

Kher-Durlabhji, N., & Lacina-Gifford, J. (1992, November). *Quest for success: Preservice teachers' views of "high-stakes" tests.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Kubiszyn, T., & Borich, G. (1990). *Educational testing and measurement: Classroom application and practice.* Glenview, IL: Scott Foresman.

Ligon, G. (1985, April). *Opportunity knocked out: Reducing cheating by teachers on student tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Perlman, C. (1985, April). *Results on a citywide testing program audit in Chicago.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Richards, T. (1989). Testmania: The school under siege. *Learning, 17*, 64-66.

Smith, M. (1991). Meaning of test preparation. *American Educational Research Journal, 28*, 521-542.

Copyright 1995 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1995, Vol. 2, No. 2, 51-57

A Typology of School Climate Reflecting Teacher Participation: A Q-technique Study

Dianne L. Taylor
Louisiana State University

Bruce Thompson
Texas A&M University and Baylor College of Medicine

Ira E. Bogotch
University of New Orleans

Understanding ways in which schools are unique is important to fostering and supporting school restructuring and improvement efforts. One way to conceptualize these differences is through a typology of school climate that reflects aspects of a school's restructuring agenda. The present paper develops such a typology by exploring teachers' participation in decision-making using Q-technique factor analysis. Three types of decision climate emerge: curricularly-focused, school-management-focused; and classroom-management-focused. This typology provides a new dimension for considering school climate, allowing the distinctiveness of individual schools relative to teacher involvement in decision-making to emerge as an important element in successful school restructuring.

Persuasive arguments are offered in the literature that effective restructuring requires empowering teachers by increasing their participation in decision-making (Carnegie Task Force on Teaching as a Profession, 1986; Prestine, 1993). Since the mid-1980s, numerous schools have experimented with empowerment in this manner. However, because schools differ, the implementation of these experiments also has differed, each reflecting characteristics unique to a given school (David, 1991).

Devaney and Sykes (1988) argue that differences in the nature of change efforts are not shortcomings of current reform undertakings; rather, they assert that building on the unique characteristics of each school is essential if school improvement and restructuring are to be successful. McLaughlin and Marsh (1990) support this assertion, citing findings from the Rand Corporation Change Agent study. According to the Rand study, the history of school improvement programs should not be disregarded; however, modifications to fit a particular school environment are almost always in order. As McLaughlin and Marsh note, "in a sense teachers and administrative staff need to 'reinvent the wheel' each time an innovation is brought into the school setting" (p. 226).

Understanding the ways in which schools differ uniquely is important to formulating policies that will foster and support restructuring efforts. One avenue for conceptualizing these differences is through a typology of school climate that reflects a school's restructuring agenda. The present paper describes development of a climate typology using data evaluating teacher participation in decision-making.

Participation and Climate

Development of the proposed typology was complicated by the vagueness of the central terms involved. For example, Lowin (1968) defines participation in decision-making as "a mode of *organizational* operations in which decisions as to activities are arrived at by the very persons who are to execute those decisions" (p. 69, his emphasis). Melcher (1976) offers a more conservative slant, noting that decision participation is "the extent to which subordinates, or other groups who are affected by decisions, are consulted with, and involved in the making of decisions" (p. 120). Implicit in Melcher's definition is the notion of degree, a concept also advanced by others (Locke & Schweiger, 1979; Vroom & Yetton, 1973). In fact, Daçhler and Wilpert (1978) propose that participation forms a continuum, ranging from exclusion to full participation "with no distinction between managers and subordinates" (p. 14).

Dianne L. Taylor is on the faculty of the Louisiana State University. Bruce Thompson is Professor of Education and Distinguished Research Fellow at Texas A & M University and Adjunct Professor of Community Medicine at Baylor College of Medicine. Ira E. Bogotch is a member of the faculty at the University of New Orleans. Correspondence regarding the paper should be addressed to Dr. Dianne L. Taylor, 111 Peabody Hall, Louisiana State University, Baton Rouge, LA 70803.

In practice, no single conception of decision participation is sufficient to account for the variety of goals, needs, and values found in schools. While one faculty may focus on decision-making efforts involving curricular matters, another may interpret participation as being oriented toward school-wide issues, while still another may center efforts on classroom teaching and student management issues. The domain of interest chosen by a faculty is likely to depend on teachers' perceptions of leadership, school needs, their own power, and shared assumptions about school priorities.

Developing a typology of climate was further complicated by the lack of clarity in the literature concerning organizational climate. According to Hoy, Tarter, and Bliss (1990), the term is "conceptually complex and vague" (p. 260). Nonetheless, Hoy et al. describe organizational climate as "a board term that refers to members' shared perceptions of the work environment" (p. 261). Like participation, climate also has been conceptualized along a continuum. According to Halpin (1966), who conducted pioneer research on school climate, the continuum ranges from open to closed.

The present study offers a somewhat more restricted view of school climate. The typology proposed here is based on teachers' reports of their participation in decision-making at schools in a reform-oriented district where policy and union contract emphasize teacher involvement in decision-making at selected schools involved in a restructuring program. A variant of factor analysis, the Q-technique described in a readable presentation by Carr (1992), was employed to develop the typology.

Method

Sample

The present study took place in a large, urban, school district in the southeast that was experimenting with teacher involvement in decision-making at some schools as part of a larger restructuring program. Among the schools in the sample were 16 schools, both elementary and high, that had been chosen as pilot sites for the district's restructuring program. Each of the selected pilot schools was paired with a non-pilot school from the same district. Matches were based on the variables of student body size, percent of students on free lunch, and organizational level. Because of the limited number of schools in the study, a non-pilot high school that does not have a match was retained in the sample. Further, one of the pilot elementary schools was dropped from the sample because of a low return rate on the questionnaire used. Thus, the final sample included 32 schools, thirty of which were matched.

All regular teachers at each school were asked to participate in the present study. Certain demographic characteristics (gender, ethnicity, and educational level) of teachers who agreed to participate were compared with school-wide profiles at each site. These comparisons are reported in Table 1. Since the profiles of the actual 637 respondents closely matched the population profiles, the samples at each school were considered reasonably representative of each school population.

Table 1
Gender, Ethnicity, and Degree of Respondents and
Total School Faculty Expressed in Percentages

	Gender		Ethnicity			Degree
ID	Female	Male	White	Black	Hispan	Masters
7	82/91	18/ 9	64/48	36/26	0/26	36/45
10	86/82	14/18	57/54	29/39	14/ 7	14/43
28	92/85	8/15	33/56	42/27	25/15	58/28
16	91/94	10/ 6	14/11	33/24	48/65	62/30
9	100/90	0/10	69/57	13/23	19/20	38/43
12	100/90	0/10	23/39	23/29	54/32	46/37
35	96/94	5/ 6	67/67	29/25	5/ 8	46/49
27	100/76	0/24	73/70	27/30	0/ 0	50/32
24	89/85	11/15	71/52	18/28	12/20	44/37
11	93/84	7/16	57/52	14/26	29/22	57/45
6	100/90	0/10	75/55	13/26	13/17	44/31
26	86/92	14/ 8	43/29	14/25	43/46	43/41
51	57/56	43/44	76/74	14/19	10/ 5	63/45
15	92/83	8/17	44/37	22/31	31/32	39/27
21	93/94	7/ 6	60/59	27/28	13/13	48/49
20	85/85	15/15	68/51	13/35	16/15	46/24
25	100/100	0/ 0	75/57	25/36	0/ 7	60/52
4	100/93	0/ 7	83/48	17/31	0/21	33/26
22	85/87	15/13	41/49	33/29	26/22	22/33
18	91/88	10/12	62/60	29/30	10/ 9	57/55
2	51/47	49/53	80/76	17/19	3/ 5	51/49
1	59/53	41/47	90/74	0/17	10/ 9	46/38
17	82/79	18/21	82/62	12/31	6/ 7	41/46
34	59/55	41/45	80/73	12/15	8/11	47/41
30	71/86	29/14	29/29	29/29	43/43	57/49
3	100/89	0/11	57/37	44/50	0/11	33/42
14	80/89	20/11	60/46	20/29	10/21	50/36
23	78/78	22/22	67/70	33/26	0/ 4	44/24
36	43/55	57/45	70/70	4/17	22/13	57/37
37	84/90	16/10	83/60	17/27	0/13	31/29
8	100/93	0/ 7	64/45	18/31	18/24	54/46
31	57/81	43/19	86/59	0/25	14/16	43/40
Total	79	21	65	19	15	47

Note: The percentage preceding the slash represents teachers responding in the study. The percentage following the slash represents all members of the faculty at the school

Instrumentation

The decision participation measure employed in the study is a subscale of a questionnaire used previously in two larger studies (Bacharach, Bamberger, Conley, &

Bauer, 1990; Bacharach, Bauer, & Shedd, 1986). Cronbach's alphas for data from this instrument are reported to range from .83 to .66 (Bacharach et al., 1990). The construct validity of the instrument has been supported as well (Taylor, Thompson, & Bogotch, 1994). The survey instrument consists of 19 items.

Results

The median score by school for each of the 19 items was computed and submitted to a Q-technique factor analysis (Gorsuch, 1983). Through the Q-technique, it is possible to isolate clusters of schools that have a similar profile of participation as reflected by responses on the questionnaire. Based on the "scree" plot of the eigenvalues associated with the factors prior to factor rotation (Thompson, 1989), three principal components were extracted and rotated to the varimax criterion. Table 2 presents the structure coefficients produced from this analysis.

Table 2
Varimax-Rotated Factor Structure Coefficients for 32 Schools

	Factor I		Factor II		Factor III			
ID	Structure	Struc^2	Structure	Struc^2	Structure	Struc^2	h^2	Secondary Variance
2	**0.893**	**79.74%**	**0.327**	**10.69%**	**0.199**	**3.96%**	**94.336%**	**14.65%**
5	**0.877**	**76.91%**	**0.308**	**9.49%**	**0.247**	**6.10%**	**92.480%**	**15.60%**
36	0.834	69.56%	0.130	1.69%	0.425	18.06%	89.194%	19.72%
1	**0.833**	**69.39%**	**0.296**	**8.76%**	**0.282**	**7.95%**	**85.996%**	**16.68%**
3	**0.823**	**67.73%**	**0.267**	**7.13%**	**0.141**	**1.99%**	**76.929%**	**9.12%**
30	0.802	64.32%	0.425	18.06%	-0.119	1.42%	83.769%	19.43%
4	0.752	56.55%	0.391	15.29%	0.357	12.74%	84.607%	28.05%
34	0.742	55.06%	0.411	16.89%	0.411	16.89%	88.817%	33.72%
14	0.741	54.91%	0.244	5.95%	0.569	32.38%	93.212%	38.36%
31	0.731	53.44%	0.072	0.52%	0.502	25.20%	79.164%	25.73%
16	0.723	52.27%	0.430	18.49%	0.134	1.80%	72.661%	20.27%
22	0.689	47.47%	0.493	24.30%	0.358	12.82%	84.510%	37.07%
37	0.649	42.12%	0.360	12.96%	0.582	33.87%	89.014%	46.90%
28	0.645	41.60%	0.328	10.76%	0.422	17.81%	70.141%	28.52%
18	0.642	41.22%	0.315	9.92%	0.566	32.04%	83.166%	41.94%
26	0.602	36.24%	0.486	23.62%	0.446	19.89%	79.682%	43.49%
20	0.577	33.27%	0.461	21.25%	0.388	15.05%	69.635%	36.36%
21	0.568	32.26%	0.520	27.04%	0.291	8.47%	67.837%	35.52%
10	**0.232**	**5.38%**	**0.784**	**61.47%**	**-0.010**	**0.01%**	**66.938%**	**5.42%**
35	**0.174**	**3.02%**	**0.779**	**60.68%**	**0.388**	**15.05%**	**78.758%**	**18.11%**
12	**0.420**	**17.64%**	**0.739**	**54.61%**	**0.151**	**2.28%**	**74.617%**	**19.89%**
25	0.557	31.02%	0.651	42.38%	0.338	11.42%	84.756%	42.43%
9	0.421	17.72%	0.639	40.83%	0.599	35.88%	94.325%	53.54%
17	-0.031	0.10%	0.633	40.07%	0.563	31.70%	71.827%	31.73%
27	0.360	12.96%	0.617	38.07%	0.404	16.32%	67.347%	29.29%
15	0.582	33.87%	0.605	36.60%	0.193	3.72%	74.171%	37.54%
24	0.522	27.25%	0.597	35.64%	0.266	7.08%	69.839%	34.24%
23	0.489	23.91%	0.543	29.48%	0.306	9.36%	62.754%	33.26%
11	**0.069**	**0.48%**	**0.158**	**2.50%**	**0.902**	**81.36%**	**84.278%**	**2.98%**
7	**0.372**	**13.84%**	**0.229**	**5.24%**	**0.689**	**47.47%**	**66.554%**	**19.04%**
8	0.543	29.48%	0.217	4.71%	0.670	44.89%	79.028%	34.18%
6	0.457	20.88%	0.385	14.82%	0.534	28.52%	64.282%	35.74%
Pre-Rotation								
Eigenvalues		12.113		7.099		6.033	25.246	
Prerotation						25.246		
Trace		21.410		2.068		1.768		
						25.246		

Note. "Secondary Variance" is variance for a school originating from factors other than the school's primary factor, e.g., for the first school listed, 10.68% + 3.97% = 14.647%. Prerotation eigenvalues and the postrotation distribution of trace are both presented (Thompson, 1989). The nine schools identified as being most prototypic are **bolded**.

BEST COPY AVAILABLE

Factor scores were computed on the three factors--one score for each of the 32 schools on each of the 19 items. Factor scores for a school typify the decision participation pattern at the school by identifying similarities and differences in school-prototypes (Kerlinger, 1986; Thompson, 1980; Thompson & Miller, 1984). Factor scores for the 19 items on each of the three prototype factors are reported in Table 3.

Table 3
Factor Scores on the 19 Items

	All Schools (n=32) Factors			Representative Schools (n=9) Factors		
Item	I	II	III	I	II	III
School to which you are assigned	0.18	0.44	0.34	0.38	0.40	-0.08
Subject/grade level(s) you are assigned to teach	0.74	1.71	-0.95	1.03	0.64	-0.23
Assignment of students to your classes	0.10	-1.16	**-1.48**	-0.44	-0.87	**-1.19**
Removing students from class for special instruction	0.41	-1.07	0.52	0.48	-0.74	-0.10
Designing or planning use of facilities	-0.46	-0.48	-0.18	-0.16	-0.56	-0.36
Budget development	-1.11	**-1.13**	0.78	-0.91	**-1.62**	1.28
Expenditure priorities	**-1.28**	-0.68	0.56	**-1.15**	-0.43	0.99
Staff hiring	-0.72	0.06	**-1.03**	-0.58	-0.13	**-1.49**
Evaluations of your performance	-0.22	**1.57**	**-2.21**	-0.41	**2.17**	**-2.09**
Student discipline codes	**-1.74**	**1.40**	1.12	**-1.67**	**1.45**	0.87
Standardized testing policy	0.05	**-1.97**	**-1.08**	-0.23	**-1.92**	**-1.37**
Grading policies	-0.24	-0.30	-0.70	-0.06	-0.49	-.59
Procedures reporting student achievement	0.03	0.56	0.51	--0.13	0.29	0.86
Student rights	**-1.14**	0.20	0.39	**-1.06**	0.06	0.53
What to teach	**2.10**	-0.29	-0.41	**2.25**	-0.30	-0.33
How to teach	**1.63**	0.83	1.10	**1.66**	0.83	0.81
Textbooks/workbooks that will be available for use	0.86	-0.32	**1.34**	0.49	0.15	**1.10**
Specific textbooks/workbooks you use in class	**1.15**	-0.20	0.99	**1.13**	-0.09	1.05
Staff development opportunities	-0.34	0.83	0.40	-0.62	1.16	0.37

Note. Factor scores are standardized to have means of zero and standard deviations of one. Scores more than one standard deviation from the mean across both analyses have been **bolded**.

As indicated by the factor scores, teachers in those schools most strongly correlated with school-prototype Factor I feel particularly involved in decisions about what to teach (e.g., factor scores of +2.1 and +2.3, respectively, for the sample of 32 schools), how to teach, and which textbooks and workbooks they use. These teachers feel that they do not participate in decisions concerning student discipline codes, spending priorities, and student rights. The first cluster of schools is characterized as *curricularly-focused*, and includes four of the five senior high schools participating in the study.

School-prototype Factor II has a profile in which teachers perceive themselves to be especially involved in decisions about the subjects and grades they are assigned to teach, their own performance evaluation, and student discipline codes. However, teachers in these schools feel particularly uninvolved in decisions regarding standardized testing policy, student assignment to class, budget development, and the removal of students for special instruction. The involvement of these teachers tended to concentrate on issues that have more of a school-wide and managerial focus, hence the second cluster is characterized as *school-management-focused*.

The smallest cluster of schools is associated with school-prototype Factor III. Teachers at these schools feel particularly involved regarding book availability, student discipline codes, and how to teach. They infrequently participate in decisions about their own performance evaluation, students' assignment to class, standardized testing policies, and staff hiring. Teachers in this third cluster tend to see the classroom as the locus of their involvement, consequently this cluster is identified as *classroom-management-focused*.

To test the invariance of these findings, two approaches were used as recommended by Thompson (1984). The first involved identification of the most representative schools for each of the three school-prototype factors based on the results reported in Table 2. The nine schools thus selected had little common variance except with their own school-prototype factor. For example, the variance in response patterns of the first representative school listed in Table 2 (school ID number 2) was common to 80% of the variance in school-prototype Factor I, while only 15% of the variance in response patterns of this school was common to Factor II (11%) or Factor III (4%). Factor scores for the nine schools are presented in Table 3. Scores greater than |1.0| across *both* the original and the invariance analyses are bolded. Factor structure coefficients for the nine schools, presented in Table 5, are similar to those for the total sample providing an upper-bound measure of confidence that the study results are stable.

Because nesting might have contributed to the stability of the structure reported above, a second invariance approach was also used. In this instance, the sample was split into two groups according to the pilot or non-pilot status of each school. Once again, responses were submitted to factor analyses and factor scores generated. Factor scores for each group are presented in Table 4. Again, scores greater than one |1.0| across both the original and the invariance analyses are bolded.

Table 4
Factor Scores on the 19 Items for the Pilot and Non-Pilot Schools

	Pilot Schools (n=15) Factors			Non-Pilot Schools (n=17) Factors		
Item	I	II	III	I	II	III
School to which you are assigned	-0.00	0.43	0.79	0.37	0.34	-0.02
Subject/grade level(s) you are assigned to teach	1.18	0.65	-0.16	-1.55	**2.49**	0.84
Assignment of students to your classes	-0.45	-0.47	**-1.70**	-0.42	-0.03	**-1.87**
Removing students from class for special instruction	0.54	-0.83	-0.26	1.35	-0.94	-0.40
Designing or planning use of facilities	-0.16	-0.57	-0.66	-0.50	-0.57	0.18
Budget development	-0.82	**-1.80**	1.38	-0.58	**-1.00**	-0.07
Expenditure priorities	**-1.23**	-0.93	1.05	-0.47	-1.14	-0.15
Staff hiring	-0.22	-0.14	**-1.22**	-0.66	-0.72	-0.51
Evaluations of your performance	-1.32	**2.40**	**-1.55**	-0.94	0.82	-0.88
Student discipline codes	**-1.27**	0.93	**1.04**	-0.66	**-1.17**	**2.42**
Standardized testing policy	-0.17	**-1.88**	**-1.18**	-0.21	-0.49	**-1.95**
Grading policies	-0.31	-0.18	-1.04	-1.11	0.21	0.13
Procedures reporting student achievement	0.12	0.63	0.67	0.27	0.09	0.33
Student rights	-0.87	0.09	0.43	-0.60	-0.89	0.65
What to teach	**2.91**	-0.25	-0.65	0.69	1.88	-0.56
How to teach	**1.05**	0.74	0.51	**1.52**	0.91	**1.27**
Textbooks/workbooks that will be available for use	0.11	0.59	0.94	2.07	-0.25	0.07
Specific textbooks/workbooks you use in class	**1.28**	-0.11	0.82	**1.47**	0.38	-0.09
Staff development opportunities	-0.37	0.69	0.79	-0.04	0.07	0.62

Note. Factor scores are standardized to have means of zero and standard deviations of one. Scores more than one standard deviation from the mean across analyses for the total sample, the pilot subsample, and the non-pilot subsample have been **bolded**.

The structure coefficients for the pilot and non-pilot schools are presented in Table 5. The factor structure for the pilot schools closely resembles that for the total sample. Among the non-pilot schools, however, seven loaded on a factor different from the one with which they were associated in the original analysis. This result suggests that data from the non-pilot schools may be somewhat unstable and should, therefore, be interpreted with caution.

Table 5
Varimax-Rotated Factor Structure Coefficients for the Nine Representative Schools, the Pilot Schools, and Non-Pilot Schools

Representative Schools (n=9)				Pilot Schools (n=15)				Non-Pilot Schools (n=17)			
	Factors				Factors				Factors		
ID	I	II	III	ID	I	II	III	ID	I	II	III
2*	.903	.313	.195	1*	.926	.217	.195	8	.851	.158	.399
5*	.895	.300	.225	34*	.848	.318	.334	31*	.846	.305	.171
1*	.878	.221	.249	20*	.738	.327	.335	14*	.821	.422	.317
3*	.863	.270	.088	28*	.715	.305	.294	36*	.775	.522	.161
				16*	.675	.494	.100	37*	.720	.426	.437
10*	.218	.827	.010	26*	.576	.554	.360	18*	.660	.406	.505
35*	.221	.787	.360					4*	.657	.568	.340
12*	.420	.776	.144	12*	.353	.811	.155				
				10*	.215	.799	.054	30	.224	.928	.077
11*	.078	.175	.924	35*	.188	.718	.501	2	.567	.769	.192
7*	.435	.115	.783	24*	.528	.644	.218	5	.605	.734	.169
				21	.540	.582	.275	15*	.328	.647	.399
								3	.563	.645	.118
				11*	.193	.063	.899	25*	.343	.638	.605
				9	.549	.526	.589	23*	.216	.615	.551
				6*	.541	.272	.589	22	.543	.610	.424
				27	.430	.480	.570				
								17*	.175	.031	.886
								7	.415	.285	.583

Note: The asterisk (*) indicates schools which load on the same factor for both the original and the invariance factor analytic computations.

Discussion and Implications

Halpin (1966) is one of the first researchers to demonstrate that each school has its own personality. Building on this concept, McLaughlin and Marsh (1970) suggest that innovations must be relevant to school personality if school improvement is to be successful. The present study focuses on a key component of that personality—that is, teacher involvement in decision making—and explores the effect of participation on patterns of school climate. A typology of school climate, involving three prototypes, is presented.

One climate type, identified as *curricularly-focused*, includes nine of the prototypic schools. The schools most associated with this factor are primarily senior high schools. Such a finding reinforces previous research that high school teachers are especially concerned with curriculum as it pertains to their subject-matter. The current emphasis on graduation requirements, which is directly linked to students' successfully demonstrating their subject-matter knowledge, understandably inclines high school teachers to focus on curricular matters. When considering the kinds of decisions in which teachers' participation is most needed, high school teachers would logically focus on issues related to the curriculum.

Decision participation extends beyond the classroom in those schools where faculty participation leans toward decisions about *school-management* issues. Although teachers in schools with this climate type are involved in decisions regarding matters that are related to the classroom, it is likely that schools fitting this prototype have a climate that is conducive to expanding teacher leadership. In schools with this climate type, teachers may be expected to extend their participation into other areas tangentially related to the classroom, such as staff hiring, staff development, and student rights.

The third climate type is labeled *classroom-management-focused*. At these schools teachers are involved in decisions that preserve their autonomy in maintaining a smoothly functioning classroom. These teachers do not view *what* to teach as a decision participation priority, rather decisions regarding *how* they teach and the standards of discipline to which students are held are of interest. At these schools, teachers appear to be concerned with managing instruction so that disciplinary disturbances are minimized. Schools with such a climate might have teachers who are relatively receptive to instructional innovations, such as cooperative learning, provided that successful behavior management strategies are also emphasized.

As the above discussion suggests, findings from the present study lend support to the contention that schools have distinct personalities. Sensitivity to this dimension of school climate will likely have positive effects on school improvement efforts. Knowing areas of strength and interest shared by teachers at a school enables reformers to tailor change programs and associated staff development to the unique outlook of a particular faculty. This study also points to a need for sensitivity to school climate on the part of policy makers as they develop restructuring agendas. Disregard for these differences could potentially compromise restructuring efforts.

References

Bacharach, S. B., Bamberger, P., Conley, S. C., & Bauer, S. (1990). The dimensionality of decision participation in educational organizations: The value of a multi-domain evaluative approach. *Educational Administration Quarterly, 26*, 126-167.

Bacharach, S. B., Bauer, S., & Shedd, J. B., (1986). *The learning workplace: The conditions and resources of teaching.* (ERIC Document Reproduction Service No. ED 279 614)

Carr, S. (1992). A primer on Q-technique factor analysis. *Measurement and Evaluation in Counseling and Development, 25*, 133-138.

Carnegie Task Force on Teaching as a Profession. (1986). *A nation prepared: Teachers for the 21st century.* Washington, DC: Carnegie Forum on Education and the Economy.

Dachler, H. P., & Wilpert, B. (1978). Conceptual dimensions and boundaries of participation in organizations: A critical evaluation. *Administrative Science Quarterly, 23*, 1-36.

David, J. (1991). What it takes to restructure education. *Educational Leadership, 48*(8), 11-15.

Devaney, K., & Sykes, G. (1988). Making a case for professionalism. In A. Lieberman (Ed.), *Building a professional culture in schools* (pp. 3-22). New York: Teachers College Press.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Halpin, W. W. (1966). *Theory and research in administration.* New York: Macmillan.

Hoy, W. K., Tarter, C. J., & Bliss, J. R. (1990). Organizational climate, school health, and effectiveness: A comparative analysis. *Educational Administration Quarterly, 26*, 260-279.

Kerlinger, F. K. (1986). *Foundations of Behavioral Research* (3rd ed.). New York: Holt, Rinehart & Winston.

Locke, E. D., & Schweiger, D. M. (1979). Participation in decision-making: One more look. In B. M. Staw (Ed.), *Research in organizational behavior* (pp. 265-339). Greenwich, CN: JAI Press.

Lowin, A. (1968). Participative decision making: A model, literature critique, and prescriptions for research. *Organizational Behavior and Human Performance, 3*, 68-106.

McLaughlin, M. W., & Marsh, D. D. (1990). Staff development and school change. In A. Lieberman (Ed.), *Schools as collaborative cultures: Creating the future now* (pp. 213-232). New York: Falmer.

Melcher, A. J. (1976). Participation: A critical review of research findings. *Human Resource Management, 15*, 12-21.

Prestine, N. (1993). Feeling the ripples, riding the waves: Making an essential school. In J. Murphy & P. Hallinger (Eds.), *Restructuring schooling: Learning from ongoing efforts* (pp. 32-62). Newbury Park, CA: Corwin.

Taylor, D., Thompson, B., & Bogotch, I. E. (1994). Investigating the construct validity of scores from a measure of teachers' participation in decision making using procrustean rotation. *Educational and Psychological Measurement, 54*, 193-198.

Thompson, B. (1980). Validity of an evaluator typology. *Educational Evaluation and Policy Analysis, 2*, 59-65.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation.* Newbury Park, CA: Sage.

Thompson, B. (1989). Prerotation and postrotation eigenvalues shouldn't be confused: A reminder. *Measurement and Evaluation in Counseling and Development, 22*(3), 114-116.

Thompson, B., & Miller, L. A. (1984). Administrators' and evaluators' perceptions of evaluation. *Educational and Psychological Research, 4*, 207-219.

Vroom, V., & Yetton, P. W. (1973). *Leadership and decision making* (pp. 11-58). Pittsburgh: University of Pittsburgh Press.

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◇◇

Please enter a subscription to Research in the Schools for:

Name: ______________________________

Institution: ______________________________

Address: ______________________________

____

		COST
Individual Subscription ($25 per year)	Number of years _______	__________
Institutional Subscription ($30 per year)	Number of years _______	__________
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years _______	__________
TOTAL COST:		__________

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
Research in the Schools
UAB School of Education
901 13th Street, South
Birmingham, AL 35294-1250

Please note that a limited number of copies of Volume 1 are available and can be purchased for the same subscription prices noted above.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

MSERA Membership/Renewal Form

(Please print or type)

NAME: ____________________

TITLE: ____________________

INSTITUTION: ____________________

MAILING ADDRESS: ____________________

____

____

PHONE: ____________________ FAX ____________________

ELECTRONIC MAIL ADDRESS: ____________________

MSERA MEMBERSHIP: New___ Renewal ___
ARE YOU A MEMBER OF AERA? Yes___ No___
WOULD YOU LIKE INFORMATION ON AERA MEMBERSHIP? Yes___ No___

DUES:	Professional	$15.00	______
	Student	$10.00	______

VOLUNTARY TAX DEDUCTIBLE CONTRIBUTION
TO MSER FOUNDATION ______

TOTAL ______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. Dorothy D. Reed (MSERA)
Headquarters, Air University
USAF, 55 LeMay Plaza South
Maxwell AFB, AL 36112-6335

RESEARCH IN THE SCHOOLS
Mid-South Educational Research Association
and The University of Alabama
Evaluation and Assessment Laboratory
Post Office Box 870231
Tuscaloosa, AL 35487-0231

NON-PROFIT ORG
U.S. POSTAGE
PAID
TUSCALOOSA, AL
PERMIT NO. 16

Soo-Back Moon
Hyosung Women's University
c/o James E McLean U of Ala
POB 870231
Tuscaloosa, AL 35487-0231

TM-CNL5

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and The University of Alabama.

Volume 3, Number 1 **Spring 1996**

Effective Teaching Behaviors for Beginning Teachers: A Multiple Perspective 1
David M. Shannon, Daniel L. Swetman, Nancy H. Barry, and John F. vonEschenbach

The Relationship of Student Attitudes Toward Science, Mathematics, English and Social Studies in U.S. Secondary Schools 13
Jianjun Wang, J. Steve Oliver, and Andrew T. Lumpe

The Verbal and Nonverbal Intelligence of American vs. French Children at Ages 6 to 16 Years 23
Alan S. Kaufman, James C. Kaufman, Nadeen L. Kaufman and Mireille Simon

The Prediction of Academic Achievement Using Non-Academic Variables 35
Susan E. Britt and Jwa K. Kim

Testing at Higher Taxonomic Levels: Are We Jeopardizing Reliability by Increasing the Emphasis on Complexity? 45
Andrea.D. Clements and Lori Rothenberg

The Selection of Female Secondary School Assistant Principals and Transformational Leadership 51
Ann Hassenpflug

A Survey of Accelerated Master of Teaching Program Graduates at The University of Memphis 61
Tiffany L. Bailey and Linda Bol

Using A Priori Versus Post-Hoc Assignment of a Concomitant Variable to Achieve Optimal Power from ANOVA, Block, and ANCOVA Designs 67
Yi-Cheng Wu and James E. McLean

308

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1996 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 3, Number 1 **Spring 1996**

EDITORS
James E. McLean, *The University of Alabama at Birmingham*
and Alan S. Kaufman, *Psychological Assessment Resources, Inc. (PAR)*

PRODUCTION EDITOR
Margaret L. Rice, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Kentucky*
Donald F. DeMoulin, *University of Memphis*
R. Tony Eichelberger, *University of Pittsburgh*
Daniel Fasko, Jr., *Morehead State University*
Ann T. Georgian, *Hattiesburg (Mississippi) High School*
Tracy Goodson-Espy, *University of North Alabama*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
JinGyu Kim, *National Board of Educational Evaluation (Korea)*
Jwa K. Kim, *Middle Tennessee State University*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Jerry G. Mathews, *Idaho State University*
Peter C. Melchers, *University of Cologne (Germany)*
Claire Meljac, *Unité de Psychopathologie de l'Adolescent (France)*
Soo-Back Moon, *Catholic University of Hyosung (Korea)*
Arnold J. Moore, *Emporia State University*
Thomas D. Oakland, *University of Florida*
William Watson Purkey, *University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Georgia Southern University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Valdosta State University*
Bruce Thompson, *Texas A & M University*

GRADUATE STUDENT EDITORIAL BOARD
Margery E. Arnold, *Texas A & M University*
Vicki Benson, *The University of Alabama*
Alan Brue, *University of Florida*
Sue E. Castleberry, *Arkansas State University*
James Ernest, *University of Alabama at Birmingham*
Robin A. Groves, *Auburn University*
Harrison D. Kane, *University of Florida*
James C. Kaufman, *Yale University*
Sadegh Nashat, *Unité de Psychopathologie de l'Adolescent (France)*
Michael D. Scraper, *Emporia State University*
Sherry Vidal, *Texas A & M University*

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 1-11

Effective Teaching Behaviors for Beginning Teachers: A Multiple Perspective

David M. Shannon, Daniel L. Swetman, and Nancy H. Barry
Auburn University

John F. vonEschenbach
East Tennessee State University

Two aspects of teaching behavior were examined in this study. Principals, teachers, college faculty, interns, and pre-interns were asked to respond to specific teaching behaviors in terms of a) how difficult they were for beginning teachers, and b) how essential they were for beginning teacher success. These behaviors were organized according to state mandated competencies for beginning teachers: a) Planning and Materials, b) Instruction and Communication, c) Classroom Organization and Environment, d) Evaluation, and e) Professional Responsibilities. Overall, the areas of Instruction and Communication and Classroom Organization and Environment were found to be most essential while the areas of Instruction and Communication and Planning and Materials were found most difficult for beginning teachers. Inconsistencies were found among the five groups of educators both in terms of how essential specific teaching behaviors were for beginning teachers and how difficult these behaviors were to demonstrate. Principals and teachers identified Classroom Organization and Environment as most essential, while college faculty, interns, and pre-interns identified the area of Instruction and Communication. While Planning and Materials was identified as the most difficult area by teachers, interns, and pre-interns, Instruction and Communication was identified by principals and college faculty. An explanation of these differences and a discussion of ways teacher education can address them are offered.

Research studies conducted to determine which teaching behaviors are most effective have varied a great deal both in terms of their design and the results reported from them. As these studies have accumulated, several reviews, summaries, and lists of teaching behaviors that are effective in producing student achievement have been compiled (Brophy & Good, 1986; Crawford & Robinson, 1983; Gage, 1978; Medley, 1977; Reynolds, 1992; Rosenshine & Stevens, 1986; Smith, 1983; Weber & Roff, 1983). Other efforts have concentrated on the identification of specific problems faced by beginning teachers (Veenman, 1984; Weinstein, 1988). Findings from these research efforts lead to the identification of specific areas of competence in which beginning teachers are expected to become proficient.

State departments of education and teacher preparation programs generally rely on such professional competencies to guide the determination of state certification standards as well as the preparation and assessment of prospective educators. These areas of competence typically represent general areas of teaching such as planning, instruction, classroom management, and evaluation. Specific indicators, usually defined in behavioral terms, are then used to further define these areas of competence so that they may be accurately observed and evaluated. Some states define more (or fewer) areas of competence and attach different labels, but the general nature of these competencies does not vary much from state to state and is consistent with research findings.

Perceptions of Effective Teaching

Although there exist commonalities among the research findings pertaining to effective teaching and the competencies defined for beginning teachers to demonstrate, there is little evidence that educators' perceptions are in agreement with these findings. Using the *Teaching Behaviors Questionnaire* (Marchant & Bowers, 1990), several studies have revealed differences among the perceptions of educators. Specifically, elementary teachers and teachers with little experience

---

David M. Shannon is Associate Professor, Department of Educational Foundations, Leadership, and Technology at Auburn University; Daniel Swetman is Assistant Professor, Department of Curriculum and Instruction, Auburn University; Nancy H. Barry is Associate Professor, Department of Curriculum and Instruction, Auburn University; and John F. vonEschenbach is Professor and Chair, Department of Curriculum and Instruction, East Tennessee State University. Please direct all correspondence to the first author at 4010 Haley Center, Auburn University, AL 36849 (334) 844-4460, FAX: (334) 844-3072, Email: SHANNDM@mail.auburn.edu.

(i.e., less than 6 years) were in higher agreement with research-based teaching behaviors than were secondary teachers and teachers with many years of experience (i.e., 25 or more years) (Marchant, 1988, 1992; Marchant & Bowers, 1988). Teachers have also been found to differ in their preference for models of teaching (Thompson, 1981). Elementary teachers placed more emphasis on models which promote the social growth of students, whereas secondary teachers preferred specific models which advance the intellectual and analytical skills of students. The research on principals' perceptions of effective teaching has also yielded inconsistent findings. While Marchant (1988, 1992) identified principals as being more accurate judges of effective teaching when compared to other educators, others have raised serious concerns about the accuracy of principals' judgements of teaching behavior in relationship to student achievement (Medley & Coker, 1987). Whether principals provide more accurate views may be questionable, but their views differ from those of teachers. Jandes, Murphy, and Sloan (1985) found that principals consistently viewed schools as more effective than did their teachers. Also, principals have rated teachers as more effective in managing student behavior than did the teachers themselves (Richardson, 1985). In both of these studies, perceptions at the elementary level were more positive than those at the secondary level. Marchant (1992) reported significant differences between the perceptions of principals and their teachers, with the greatest discrepancy between secondary principals and secondary teachers.

Research on preservice teachers' percepfions of effective teaching suggest a developmental pattern. Students enrolled in their first or second year of teacher education programs have generally been concerned with instructional strategies which are more teacher-centered and define an effective teacher as one who is caring, compassionate, friendly, and patient (Placek & Dodds, 1988; Turley, 1994; Weinstein, 1989; Wilson & Cameron, 1994). These perceptions of effective teaching become more concerned with student involvement and student learning as preservice teachers near the completion of student teaching (Killen, 1994; Placek & Dodds, 1988; Wilson & Cameron, 1994). The perceptions of education students and their faculty advisors have generally been identified as being less accurate than those of both teachers and principals (Marchant, 1988; 1992).

These inconsistencies between teacher educators and practitioners (i.e., teachers and principals) create discrepancies between what is emphasized in teacher education programs and the demands of the real world. Therefore, beginning teachers often enter teaching with unrealistic optimism about what to expect and are likely to have little resilience to failure (Weinstein, 1988). When beginning teachers come directly from the college classroom into the public schools, they often experience "reality shock" which can be overwhelming (Veenman, 1984). This shock may lead student teachers to abandon what they have learned and to embrace the practical "survival" advice offered by their cooperating teacher, creating a gap between the "theory" learned in methods classes and the "practice" adopted in the classroom.

Part of the reason for this gap is a difference in perspective. Beginning teachers generally draw from their academic background. Therefore, they base most of their expectations and beliefs about teaching on coursework, field experiences, and student teaching from their teacher preparation programs. Unfortunately, field experiences, including student teaching are often limited (Cochran, DeRuiter, & King, 1993; Griffen, 1989; Shannon, 1994). On the other hand, principals and veteran teachers draw more from their experiential background. They have firsthand experiences of being beginning teachers and have observed closely the struggles of other beginning teachers in the field. Principals are expected to observe and assess the effectiveness of their teaching staff, and teachers are expected to implement and demonstrate these competencies. However, these educators do not often play a major role in teacher education programs. Reactions and input from these educators concerning competencies for beginning teachers would be very informative and would help to address the gap which exists between theory and practice.

Purpose

The purpose of this study was to identify which specific components of effective teaching are most essential and most difficult for beginning teachers in order to provide useful feedback and guidance for teacher education programs. Perceptions were gathered from teachers, principals, teacher educators, interns/student teachers, and pre-intern students. It is essential that information regarding beginning teacher behaviors be gathered from educators engaged in both the theory (i.e., teacher educators, interns, and pre-interns) and the practice (i.e., principals and teachers) of teaching for several reasons. First, the identification of specific behaviors which are perceived as important and/or problematic for beginning teachers provides a basis for examining the content of teacher education programs. The teacher education curriculum should reflect those behaviors identified as most essential and begin to concentrate on those identified as problematic. Second, any discrepancies between theory (i.e., teacher educators, interns, and pre-interns) and practice (i.e., principals and teachers) provide useful feedback to teacher education programs regarding the extent to which they are equipping graduates with the knowledge and skills necessary to meet the demands of the "real world" of teaching. What knowledge and skills are perceived to be most essential for teacher survival in this "real world" and how difficult is it for teachers to learn these skills? These are serious issues to be explored as teacher education programs continue to prepare teachers for the profession of teaching.

Methods

Sample

The sample for this study was drawn from two primary sources: public K-12 schools listed in the State Education Directory and a large southeastern land-grant university. The demographic information for each sample is summarized in Table 1.

Public School Systems Sample. Fifty school districts were randomly sampled from the statewide directory of public schools in Alabama. From each of these school districts, three schools (1 elementary, 1 middle, 1 high school) were randomly selected. The principal of each school was sent a packet containing a survey instrument for an instructional leader and three surveys for teachers. The principal was asked to identify the instructional leader (i.e., principal or assistant principal) most responsible for the evaluation of teachers and to distribute the survey to that person. The principal was also asked to distribute the three teacher surveys to teachers with varying degrees of experience. This process resulted in the sampling of 150 instructional leaders, and 450 teachers across the state.

Table 1 •
Demographic Background of Samples

	Principals (n=52)	Teachers (n=108)	College Faculty (n=16)	Pre-Interns (n=93)	Interns (n=292)
Gender					
Male	34 (65%)	19 (18%)	11 (69%)	11 (12%)	41 (14%)
Female	17 (33%)	88 (81%)	5 (31%)	77 (85%)	242 (83%)
Missing	1 (2%)	1 (1%)	0 (0%)	5 (5%)	9 (3%)
Highest Degree					
Bachelors	2 (4%)	35 (32%)			
Masters	25 (48%)	60 (56%)	1 (6%)		
Specialist	6 (12%)	5 (5%)			
Doctorate	7 (14%)	1 (1%)	13 (81%)		
Admin. Assist	11 (21%)	7 (7%)			
Missing	1 (2%)	0 (0%)	2 (13%)		
Area of Certification					
Early Childhood Education	1 (2%)	13 (12%)	0 (0%)	11 (12%)	55 (19%)
Elementary Education	15 (29%)	53 (49%)	3 (19%)	15 (16%)	80 (27%)
Secondary Education	27 (52%)	46 (43%)	10 (63%)	53 (57%)	97 (33%)
Special Education	3 (6%)	6 (6%)	4 (25%)	5 (5%)	35 (12%)
Vocational Education	1 (2%)	1 (1%)	4 (25%)	0 (0%)	13 (5%)
Other	9 (17%)	11 (10%)	2 (13%)	9 (10%)	16 (6%)
Teaching Experience					
None	0 (0%)	0 (0%)			
Grades PK-3	10 (19%)	43 (40%)			
Grades 4-6	16 (31%)	42 (39%)			
Grades 7-12	35 (67%)	60 (56%)			
College	4 (8%)	8 (7%)			

BEST COPY AVAILABLE

A total of 52 instructional leaders (35%) and 108 teachers (24%) returned completed survey instruments. Of the 52 instructional leaders, 40% were from elementary schools, 29% from middle schools, and 31% from high schools. The greatest percentage stated that they were working in rural school settings (50%), followed by urban (31%), and suburban (15%) settings. Demographically, teachers were very similar to principals as 37 percent were teaching in elementary schools, 33% in middle schools, and 30% in high schools. Forty-one (41) percent of these teachers said they were teaching in rural school settings, 31% in urban settings, and 23% in suburban settings. The majority (65%) of principals were male, whereas 81% of the teachers were female. Since principals must generally hold a master's degree in order to become certified, it is not surprising that a greater percentage of principals had completed education beyond the bachelors degree.

Higher Education Sample. The samples from higher education consisted of teacher educators, interns, and pre-interns. All faculty members who were directly involved in the supervision of interns (n=28) were asked to complete the survey instrument and distribute copies of the pre-intern instrument to education students currently enrolled in their classes.

Sixteen of the 28 (57%) sampled teacher education faculty participated in the study. The sampling of pre-service teacher education classes resulted in a total of 93 pre-interns. All students who completed their internship during the fall quarter of 1992 and the spring quarter of 1993 were included in the intern sample, resulting in a total of 292 interns (91 in the fall, 201 in the spring). The majority of the college faculty and pre-interns samples were associated with secondary certification areas. A higher percentage of interns were enrolled in early childhood or elementary certification programs. The sample of intern students is more representative of students enrolled in teacher education programs within the college.

Instrumentation

Each subject received a survey instrument soliciting information on perceived difficulty and importance of specific teaching behaviors for beginning teachers. This information was organized under five headings consistent with the state-mandated competencies: a) *Planning and Materials*, b) *Instruction and Communication*, c) *Classroom Organization and Environment*, d) *Evaluation*, and e) *Professional Responsibilities* (Alabama State Department of Education, 1989). These areas of teaching behavior are further defined in Figure 1.

Planning and Materials
- selects and states long range goals and short-term measurable objectives
- identifies instructional strategies
- prepares instructional materials
- selects evaluation strategies
- matches materials with objectives
- uses materials to accommodate student differences
- uses curriculum guides of courses of study for planning instruction

Instruction and Communication
- maintains a high level of student involvement
- monitors student comprehension and reteaches material as needed
- asks high-level thinking questions throughout lesson
- provides clear goals and objectives in a logical and organized sequence
- provides opportunities for all students to be actively involved

Classroom Management and Environment
- begins lesson promptly
- creates smooth transitions between activities
- maintains on-task behavior
- enforces discipline rules consistently
- teaches management routines and rules

Evaluation
- states the criteria for student evaluation
- establishes and maintains standards and time-lines for student assignments
- provides systematic feedback on student work
- uses multiple means of assessment
- modifies assessment to accommodate different learners
- uses assessment results to revise instruction

Professional Responsibilities
- uses ideas from formal coursework, workshops, inservices, etc.
- shares ideas, materials, resources with peers
- assists in development of school curriculum
- participates in professional organizations
- relates effectively with colleagues, principals, supervisors, and other support staff

Figure 1 - Indicators of Teaching Behaviors

Under each heading, specific behaviors were described and the subject was asked to respond first by indicating how essential each behavior is to effective teaching, and second, how difficult the behavior is for beginning teachers to learn. Each subject was asked to respond along a 4-point scale for each item. The labels of "not essential," "somewhat essential," "essential," and "very essential" were attached to the 4-point scales used to evaluate how essential each teaching behavior was for beginning teachers. The word "difficult" was substituted for "essential" for each of the difficulty scales. The specific behaviors listed were based upon the state indicators of teacher competence and supported by research findings (Brophy & Good, 1986; Doyle, 1986; Reynolds, 1992; Rosenshine & Stevens, 1986; Veenman, 1984; Weinstein, 1988; 1989).

The reliability for each of the ten subscales was estimated using Cronbach's coefficient alpha (α). These reliability estimates are summarized in Table 2. The internal consistency estimates for the five essential behavior subscales ranged from .72 to .84 (median=.79), with a total essential behavior scale reliability of .93. Estimates of reliability for the five difficulty scales ranged from .73 to .84, with a median of .80. The total difficulty scale reliability was estimated at .94.

Table 2
Summary of Reliabilities for Essential and Difficulty Scales

Scale	Number of Items	Alpha (α)
Essential Scales		
Planning and Materials	9	.72
Classroom Organization and Environment	8	.73
Instruction and Communication	9	.84
Evaluation	8	.79
Professionalism	6	.80
Total Scale	40	.93
Difficulty Scales		
Planning and Materials	9	.77
Classroom Organization and Environment	8	.80
Instruction and Communication	9	.84
Evaluation	8	.83
Professionalism	6	.73
Total Scale	40	.94

Analysis of Data

Each participant responded to specific teaching behaviors which were listed under one of the five headings described above. A value between 1 and 4 was assigned to each response. These values were then used to compute an average for each subject. This procedure was repeated for each area of teaching behavior, resulting in a total of ten scores (i.e., five regarding how essential teaching behaviors were and five regarding the difficulty they presented for beginning teachers).

A mixed-model repeated measures ANOVA design was applied. This analysis was performed to explore overall differences in perceptions across the five specified areas of teaching behaviors among the five groups of educators. Post-hoc Tukey tests were then applied to detect specific pairwise differences.

Results

The results from the mixed-model ANOVAs are summarized in Table 3. The results from these analyses revealed statistically significant differences among the five groups of educators both in terms of how essential (F=2.65, $p < .05$) and how difficult (F=8.22, $p < .001$) teaching behaviors are for beginning teachers. Means and standard deviations for each of the five groups are reported in Table 4. In general, preservice teachers (interns and pre-interns) perceived all teaching behaviors as more essential, but less difficult than did the other groups of educators.

Table 3
Summary of Mixed-Model Results

Essential Behaviors	*SS*	*df*	*MS*	*F*
Between Subjects				
Group (A)	6.42	4	1.60	2.65*
Subjects/Group (S/A)	337.04	556	.61	
Within Subjects				
Behaviors (B)	20.78	4	5.19	65.59**
Group X Behaviors (AB)	7.04	16	.44	5.55**
Behaviors X Sub/Groups (B X S/A)	176.14	2224	.08	

Difficult Behaviors	*SS*	*df*	*MS*	*F*
Between Subjects				
Group (A)	29.88	4	7.47	8.22**
Subjects/Group (S/A)	505.01	556	.91	
Within Subjects				
Behaviors (B)	48.46	4	12.11	117.80**
Group X Behaviors (AB)	7.79	16	.49	4.74**
Behaviors X Sub/Groups (B X S/A)	228.71	2224	.10	

*$p < .05$
**$p < .001$

BEST COPY AVAILABLE

Table 4
Group Means and Standard Deviations on Perceptions of Essential and Difficult Teaching Behaviors

	Principals (n=52) Mean (SD)	Teachers (n=108) Mean (SD)	College Faculty (n=16) Mean (SD)	Pre-Interns (n=93) Mean (SD)	Interns (n=292) Mean (SD)	**TOTAL SAMPLE (N=561) Mean (SD)**
Essential Behavior Scales (Group Means)	**3.32**	**3.30**	**3.278**	**3.34**	**3.416**	
Planning and Materials	3.24 (.46)	3.21 (.41)	3.13 (.38)	3.21 (.46)	3.28 (.41)	**3.25 (.42)**
Classroom Organization and Environment	3.47 (.43)	3.51 (.36)	3.40 (.29)	3.42 (.44)	3.54 (.35)	**3.50 (.38)**
Instruction and Communication	3.45 (.43)	3.44 (.44)	3.44 (.36)	3.58 (.40)	3.57 (.39)	**3.52 (.41)**
Evaluation	3.32 (.48)	3.36 (.46)	3.37 (.34)	3.27 (.47)	3.33 (.42)	**3.32 (.44)**
Professionalism	3.13 (.56)	3.06 (.51)	3.05 (.43)	3.24 (.52)	3.36 (.48)	**3.25 (.51)**
Difficulty Behavior Scales (Group Means)	**2.04**	**2.042**	**2.346**	**1.97**	**1.868**	
Planning and Materials	2.29 (.46)	2.22 (.52)	2.53 (.53)	2.13 (.47)	2.12 (.49)	**2.17 (.50)**
Classroom Organization and Environment	1.95 (.41)	2.09 (.57)	2.30 (.63)	1.99 (.52)	1.86 (.50)	**1.95 (.53)**
Instruction and Communication	2.30 (.52)	2.21 (.57)	2.65 (.52)	2.03 (.55)	1.93 (.53)	**2.05 (.57)**
Evaluation	2.06 (.61)	2.06 (.58)	2.34 (.50)	1.95 (.57)	1.87 (.51)	**1.95 (.55)**
Professionalism	1.60 (.46)	1.63 (.46)	1.91 (.46)	1.75 (.54)	1.56 (.46)	**1.62 (.48)**

The results also revealed differences among the five teaching areas, both in terms of being essential and difficult. Overall, the two behavior areas: (a) *Instruction and Communication*, and (b) *Classroom Organization and Environment*, were identified as being more essential (F=65.59, p < .001) than the remaining three areas of teaching. The teaching areas of (a) *Planning and Materials*, and (b) *Instruction and Communication*, were evaluated as being the most difficult for beginning teachers (F=117.80, p <. 001).

These results were, however, qualified by the interactions between group and behaviors, both in terms of how essential (F=5.55, p < .001), and how difficult (F=4.74, p < .001) these behaviors are for beginning teachers. All five groups agreed that the areas of *Instruction and Communication*, and *Classroom Organization and Environment* were the two most essential areas of teaching. While principals and teachers identified *Classroom Organization and Environment* as most essential, the remaining groups (college faculty, pre-interns, and interns) chose *Instruction and Communication*. The two least essential areas identified were *Planning and Materials* and *Professionalism*. However, pre-interns and interns identified *Planning and Materials* as least essential, while principals, teachers, and college faculty perceived *Professionalism* as least essential.

All five groups agreed that the area of *Professionalism* was least difficult. While *Planning and Materials* was evaluated as one of the least essential areas, it was identified as the most difficult area by teachers, pre-interns, and interns. The area of *Instruction and Communication* was identified as most difficult by principals and college faculty.

Discussion

What behaviors are most essential for beginning teachers and which ones are most difficult to demonstrate? The results from the current study suggest that *Instruction and Communication* and *Classroom Organization and Environment* were the two most essential categories of behaviors for beginning teachers to be effective. In terms of difficulty, *Instruction and Communication* and *Planning and Materials* were identified as the two most difficult areas of teaching behaviors. These findings are consistent with those of Veenman (1984) and Weinstein (1988), in which instructional and planning behaviors were identified among the most problematic for beginning teachers. Adams (1982) and Weinstein (1989) also found that preservice teachers expressed great concern for instructional and communication behaviors.

Despite this initial agreement, many inconsistencies were revealed when examining specific teaching behaviors. In general, the groups of preservice teachers (i.e., interns and pre-interns) tended to evaluate teaching behaviors as essential for beginning teachers but not very difficult to demonstrate. On the other hand, the other groups of educators (i.e., college faculty, principals, and teachers) tended to evaluate behaviors as somewhat less essential, but more difficult. We offer two explanations for these inconsistencies: experience and communication.

The first issue has to do with classroom teaching experience. Preservice teachers in this study are very confident in the ability of beginning teachers and judge teaching behaviors to present them with little difficulty. They are, however, basing their judgement of such difficulty on a very limited amount of experience. This inexperience makes them very susceptible to what Weinstein (1988) identified as "unrealistic optimism" which can often lead to "the collapse of the missionary ideals formed during teacher training by the harsh and rude reality of everyday classroom life" (Veenman, 1984, p.143).

Many teacher education programs, unlike other professional preparation programs, are limited in terms of the field-based experiences they offer preservice teachers. Limiting these experiences contributes to the unrealistic expectations held by preservice teachers. As teachers gain more experience, both in teaching and supervising teachers, their perceptions become more sophisticated and begin to better represent the complex nature of the teaching process. This is consistent with Fuller's theory that perceptions of teaching progress along a developmental continuum as teachers gain more experience (Fuller, 1969). A parallel can also be made to the expert-novice teacher literature which reports beginning (novice) teachers as having a much more narrow view of teaching (Berliner, 1986; Calderhead, 1981).

The second explanation is that of communication, both between practitioners and teacher educators and between teacher educators and students. The voice of the practitioner is too often excluded in teacher education. The perspectives offered by practicing teachers and principals provide preservice teachers the opportunity to capture a more realistic view of the complex, contextual nature of teaching before they begin to teach (or student-teach). Providing preservice teachers with a view of teaching from multiple perspectives helps to prevent the development of narrowly focused teachers with unrealistic expectations and also serves to minimize the effects of "reality shock" once they begin their first year of teaching. The practitioner's perspective must be communicated to teacher educators.

Communication must also be clear within teacher education programs. The perceptions of college faculty must be communicated effectively to preservice teachers and the needs of preservice teachers communicated effectively to college faculty. The amount of emphasis instructors place on a group of teaching behaviors will somewhat depend upon how essential and difficult they perceive them to be. Similarly, the amount of effort students exert to learn these behaviors will also depend on their perception of their importance and difficulty (Lanier & Little, 1986). What students are learning in their preservice programs may be very different from the objectives established by teacher educators for their teacher preparation program. This incongruence between student learning and program objectives is supported by other researchers (Feiman-Nesmer & Buchmann, 1985, 1986). Teacher education would become more efficient if both the teacher and learner are in agreement with, or at least understand, each other's perceptions as to the importance and difficulty of teaching behavior.

Recommendations

A gap between theory and practice remains despite teacher education's continual efforts to address this gap. Addressing the issues of experience and communication may help to narrow this gap. The issue of experience can be addressed by simply providing preservice teachers with such experience during their teacher preparation program. Cochran, DeRuiter, and King (1993) described the knowledge acquired by preservice teachers as *pedagogical content knowing* (PCKg) and outline a model for its development. This model includes the development of knowledge about subject matter, pedagogy, students, and environmental contexts resulting from a variety of clinical experiences. These experiences allow preservice teachers multiple opportunities to observe other teachers, teach on their own, receive feedback, and reflect upon their teaching.

The scheduling of professional coursework in "blocks" should continue. This structure requires faculty from various methods courses to plan course requirements and field experiences jointly and afford preservice teachers greater opportunities to gain the professional knowledge and skill they will need to become effective teachers (Shannon, 1994). In addition, students completing courses in these "blocks" are given immediate opportunities to practice what they learned in courses during field experiences proceeding their internships.

The issue of communication between practitioners and teacher education programs is best addressed by increasing the involvement of practitioners, either directly or indirectly. This helps to bridge the gap that commonly exists between the theory and practice which shapes the definition of effective teaching. We support further use of approaches such as case-based instruction (Kleinfield, 1991; Merseth, 1990; Shulman, 1986; Sykes, 1989; Welty, 1989). Capturing teaching from different contexts in the form of written or video "cases" and discussing these in methods classes provides an indirect link between preservice teachers and practicing teachers. The discussion among preservice teachers and college faculty, which results from such cases, also affords preservice teachers the opportunities to apply what they have learned in coursework to the practice of teaching. Many resources are available to teacher educators which contain "cases" written by both teachers and university faculty members (Greenwood & Parkay, 1989; Kowalski, Weaver, & Henson, 1990; Shulman & Colbert, 1988; Shulman & Colbert, 1987; Silverman, Welty, & Lyon, 1992).

Preservice teachers should also be provided with greater opportunities to engage in reflective thinking to improve not only their perceptions of effective teaching, but their actual teaching ability. Teacher educators play a critical role in fostering preservice teachers' reflective thinking skills. Many approaches, such as reflection journals, microteaching, audio and video-taped teaching, and other self-assessment techniques, should continue to be used in an effort to foster reflective thinking (Colton and Sparks-Langer, 1993; Cruickshank & Metcalf, 1993; Hatton & Smith, 1994; Sparks-Langer & Colton, 1991; Sparks-Langer, Simmons, Pasch, Colton, & Starko, 1990). For an annotated bibliography on preservice teacher reflection, see Stewart (1994).

Another vehicle of reflection which has continued to grow within teacher education programs is that of portfolios (Barton & Collins, 1993; Bird, 1990; Cole 1992; Collins, 1990; Ryan & Kuhs, 1993; Shannon, Ash, Barry, & Dunn, 1995; Smolen & Newman, 1992; Wolf, 1991a, 1991b). Portfolios provide many opportunities for professional growth as preservice teachers reflect upon their teaching in order to determine what evidence best supports their ability to teach. An important feature of using portfolios is that of collaboration. As preservice teachers gain additional opportunities to interact with colleagues (e.g., peers, cooperating teachers, university supervisors), they gain valuable experience and feedback which will allow them to improve their teaching abilities.

A more direct link can be established through the involvement of practicing teachers as "clinical instructors" within teacher education. These clinical instructors serve a direct role in the instruction of preservice teachers, a role well beyond that of a cooperating teacher during the student-teaching experience. Establishing Professional Development Schools (PDS) between universities and local schools promotes opportunities for preservice teachers to work more closely with these clinical instructors (Kunkel et al., 1992). For those interested in further information regarding PDS, Abdal-Haqq (1992) provides an annotated bibliography summarizing issues related to PDS while Teitel (1994) provides a discussion of the issues faced in establishing and institutionalizing such partnerships.

Conclusion

The perceptions which preservice teachers have upon admittance to teacher education programs have a very strong influence on how teacher candidates perform at the end of their program (Kennedy, 1991). It is important then that preservice teachers develop realistic expectations about their future teaching careers so that they are more likely to succeed and remain in the profession. As preservice teachers prepare for teaching under more realistic conditions, the tendency to believe that "it won't happen to me" is reduced. Thus, teacher education programs need to provide more opportunities for communication between practitioners and teacher educators and engage preservice teachers in reflective experiences that help to promote their professional growth. Under these conditions, preservice teachers can explore and analyze the demands of teaching (e.g., planning, instruction, classroom management, and evaluation) so that they can refine their expectations about teaching and be better prepared to face the challenges that await them.

The comparisons made among educators in this study have indicated that the beliefs of principals and their teachers are more closely aligned than those between teacher educators and their students. Although these group comparisons were made with a small sample of teacher educators and should be confirmed with additional data, we believe that they convey a reason to be concerned. Teacher preparation programs exist to equip preservice teachers with the knowledge and skills necessary to become effective teachers. As teacher educators and preservice teachers work together to better understand each others' conceptual orientation toward teaching, they help to facilitate the execution of teacher education's mission.

References

Abdal-Haqq, I. (1992). Professional Development Schools: An annotated bibliography of selected ERIC resources. *Journal of Teacher Education, 43*(1), 42-45.

Adams, R. D. (1982). Teacher development: A look at change in teacher perceptions and behavior over time. *Journal of Teacher Education, 33*(4), 40-43.

Alabama State Department of Education (1989). *Criteria for the Alabama Professional Education Personnel Evaluation System*. Montgomery, AL: Author.

Barton, J. & Collins, A. (1993). Portfolios in teacher education. *Journal of Teacher Education, 44*(3), 200-210.

Berliner, D. C. (1986). In search of expert pedagogue. *Educational Researcher, 17*(7), 5-13.

Bird, T. (1990). The schoolteacher's portfolio: An essay of possibilities. In J. Millman and L. Darling-Hammond (Eds.), *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers* (2nd ed., pp. 241-256). Newbury Park, CA: Sage Publications.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of Research on Teaching*, (3rd ed.). New York: Macmillan.

Calderhead, J. (1981). A psychological approach to research on teacher's decision making. *British Educational Research Journal, 7*, 51-57.

Carnegie Commission. (1986). *A nation prepared: Teachers for the 21st century*. New York: Carnegie Forum on Education and the Economy.

Cochran, K. E., DeRuiter, J. A., & King, R. A. (1993). Pedagogical content knowing: An integrative model for teacher preparation. *Journal of Teacher Education, 44*(4), 263-272.

Cole, D. J. (1992, February). *The developing professional: Process and product portfolios*. Paper presented at the annual meeting of the American Association of Colleges of Teacher Education (AACTE), San Antonio, TX.

Collins, A. (1990, April). *Transforming the assessment of teachers: Notes on a theory of assessment for the 21st century*. Paper presented at the annual meeting of the National Catholic Education Association, Toronto, CA. (ERIC Document Reproduction No. ED 321 362)

Colton, A. B., & Sparks-Langer, G. M. (1993). A conceptual framework to guide the development of teacher reflection and decision making. *Journal of Teacher Education, 44*(1), 45-54.

Crawford, J., & Robinson, C. (1983). A review of empirical research in classroom management. In W. Weber, L. Roff, J. Crawford & C. Robinson (Eds.), *Classroom management: Reviews of the teacher education and research literature* (pp. 43-62). Princeton, NJ: Educational Testing Service.

Cruickshank, D. R., & Metcalf, K. K. (1993). Improving preservice teacher assessment through on-campus laboratory experiences. *Theory into Practice, 32*(2), 86-92.

Doyle, W. (1986). Classroom organization and management. In M. Wittrock (Ed.), *Handbook of research on teaching*, (3rd ed.), New York: Macmillan.

Feiman-Nesmer, S., & Buchmann, M. (1985). *The first year of teacher preparation: Transition to pedagogical thinking?* (Research Series No. 156). East Lansing, MI: The Institute for Research on Teaching, Michigan State University. (ERIC Document Reproduction Service No. ED 256 753).

Feiman-Nesmer, S., & Buchmann, M. (1986). *Knowing, thinking and doing in learning to teach: A research framework and some initial results*. (Report No. SP028140). East Lansing, MI: The Institute for Research on Teaching, Michigan State University. (ERIC Document Reproduction Service No. ED 274 653).

Fuller, F. F. (1969). Concerns for teachers: A developmental perspective. *American Educational Research Journal, 6*, 207-226.

Gage, N. L. (1978). *The scientific basis of the art of teaching*. New York: Teachers College Press.

Greenwood, G. E., & Parkay, F. W. (1989). *Case studies for teacher decision making*. New York: Random House.

Griffen, G. A. (1989). A descriptive study of student teaching. *The Elementary School Journal, 89*(3), 343-364.

Hatton, N., & Smith, D. (1994, July). *Facilitating reflection: Issues and research*. Paper presented at the annual meeting of the Australian Teacher Education Association, Brisbane, Queensland, Australia. (ERIC Documant Reproduction Service No. 375 110).

Hoffman, D. E., & Roper, S. S. (1985, April). *How valuable is teacher training to beginning teachers?* Paper presented at annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 258 967).

Jandes, H. L., Murphy, J. F., & Sloan, C. A. (1985). The effective school research and Illinois public schools: A mismatch? *Illinois School Research and Development, 21*(3), 16-24.

Kennedy, M. (1991). Some surprising findings on how teachers learn to teach. *Educational Leadership, 49*(3), 14-17.

Killen, R. (1994, July). *Student teachers' perceptions of successful and unsuccessful events during practice teaching*. Paper presented at the annual meeting of the

Australian Teacher Education Association, Brisbane, Queensland, Australia. (ERIC Documant Reproduction Service No. 375 109).

Kleinfield, J. (1991, April). *The case method in teacher education: Effects on preservice teachers.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kowalski, T. J., Weaver, R. A., & Henson, K. T. (1990). *Case studies on teaching.* New York: Longman.

Kunkel, R., Richardson, E., Halpin, G., Garret, F., Tomlin, J., Swinney, A, & Pruet, S. (1992, February). *Public school and university collaboration: Steps to take for success: The Auburn model.* Paper presented at the annual meeting of the Association of Teacher Educators, Orlando, FL.

Lanier, J. E., & Little, J. W. (1986). Research on teacher education. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillian.

Marchant, G. J. (1992). Attitudes toward research-based effective teaching behaviors. *Journal of Instructional Psychology, 19*(2), 119-126.

Marchant, G. J. (1988, October). *Attitudes toward research-based effective teaching behaviors from teachers, principals, and college faculty and students.* Paper presented at the annual meeting of the Mid-Western Educational Research Association. (ERIC Document Reproduction Servive No. ED 303 449.)

Marchant, G. J., & Bowers, N. D. (1990). An attitude inventory for research-based effective teaching behaviors. *Educational and Psychological Measurement, 50,* 167-174.

Marchant, G. J., & Bowers, N. D. (1988, April). *Teacher agreement with research-based effective teaching behaviors.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Medley, D. M. (1977). *Teacher competency and teacher effectiveness: A review of the process-product research.* Washington, DC: American Association of College for Teacher Education.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgements of teacher performance. *Journal of Educational Research, 80,* 242-247.

Merseth, K. K. (1990, June). *The case for cases in teacher education.* Unpublished manuscript, Washington, DC: AAHE.

National Commission on Excellence in Education. (1983). *A nation at risk.* Washington, DC: United States Department of Education.

Placek, J. H., & Dodds, P. (1988). A critical incident study of preservice teacher beliefs about teaching success and non-success. *Research Quarterly for Exercise and Sport, 59*(4), 351-358.

Reynolds, A. (1992). What is competent beginning teaching? A review of the literature. *Review of Educational Research, 62*(1), 1-35.

Richardson, M. S. (1985). Perceptions of principals and teachers on effective management of student behavior. *Spectrum, 3*(3), 25-30.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. Wittrock (Ed.), *Handbook of research on teaching.* (3rd ed.). New York: Macmillan.

Ryan, J. M., & Kuhs, T. M. (1993). Assessment of preservice teachers and the use of portfolios. *Theory into Practice, 32*(2), 75-81.

Shannon, D. M. (1994). The development of preservice teacher knowledge. *The Professional Educator, 17*(1), 31-39.

Shannon, D. M., Ash, B., Barry, N., & Dunn, C. (1995, April). *Implementing a portfolio-based evaluation system for preservice teachers.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Shulman, L. (1986)... Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Shulman, J. H., & Colbert, J. A. (1988). *The intern teacher casebook.* Eugene, OR: ERIC Clearinghouse on Educational Management.

Shulman, J. H., & Colbert, J. A. (1987). *The mentor teacher casebook.* Eugene, OR: ERIC Clearinghouse on Educational Management.

Silverman, R., Welty, W. M., & Lyon, S. (1992). *Case studies for teacher problem solving.* New York: McGraw-Hill.

Smith, D. C. (Ed.). (1983). *Essential knowledge for beginning educators.* Washington, DC: American Association of Colleges for Teacher Education. (ERIC Document Reproduction Service No. SP 022 600).

Smolen, L. & Newman, C. (1992, February). *Portfolios: An estimate of their validity and practicality.* Paper presented at the annual meeting of the Eastern Educational Research Association, Hilton Head, SC.

Sparks-Langer, G. M., & Colton, A. B. (1991). Synthesis of research on teachers' reflective thinking. *Educational Leadership, 48*(6), 37-44.

Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., & Starko, A. (1990). Reflective pedagogical thinking: How can we promote it and measure it? *Journal of Teacher Education, 41*(4), 23-32.

Stewart, D. K. (1994). Reflective teaching in preservice teacher education: An annonated bibliography from the ERIC database. *Journal of Teacher Education, 45*(4), 298-302.

Sykes, G. (1989). Learning to teach with cases. *Colloquy, 2*(2), 7-13.

Teitel, L. (1994). Can school-university partnerships lead to the simultaneous renewal of schools and teacher education? *Journal of Teacher Education, 45*(4), 245-252.

Thompson, B. (1981). Teachers' preferences for various teaching models. *NASSP Bulletin, 65*(446), 96-100.

Turley, S. (1994, April). *"The way teachers teach is, like, totally whacked": The student voice on classroom teachers.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Documant Reproduction Service No. ED 376 164).

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*(2), 143-178.

Weber, W. A. & Roff, L. A. (1983). A review of the teacher education literature on classroom management. In W. Weber, L. Roff, J. Crawford & C. Robinson (Eds.), *Classroom management: Reviews of the teacher education and research literature* (pp. 7-42). Princeton, NJ: Educational Testing Service.

Weinstein, C. S. (1988). Preservice teachers' expectations about the first year of teaching. *Teaching and Teacher Education, 4*(1), 31-40.

Weinstein, C. S. (1989). Teacher education students' preconceptions of teaching. *Teaching and Teacher Education, 4*(2), 53-60.

Welty, W. (1989, July/August). Discussion method teaching. *Change*, 41-49.

Wilson, S., & Cameron, R. (1994, July). *What do student teachers perceive as effective teaching?* Paper presented at the annual meeting of the Australian Teacher Education Association, Brisbane, Queensland, Australia. (ERIC Document Reproduction Service No. ED 375 108).

Wolf, K. (1991a). The schoolteacher's portfolio: Issues in design, implementation, and evaluation. *Phi Delta Kappan, (73)*, 129-136.

Wolf, K. (1991b). *Teaching portfolios: Synthesis of research and annotated bibliography.* San Francisco, CA: Far West Laboratory.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 13-21

The Relationship of Student Attitudes Toward Science, Mathematics, English and Social Studies in U.S. Secondary Schools*

Jianjun Wang
California State University

J. Steve Oliver
University of Georgia

Andrew T. Lumpe
University of Toledo

Much of the curriculum in the U.S. secondary schools is filled by courses in science, mathematics, language arts, and social studies to facilitate students' growth in the four major academic areas. The interrelationships of student attitudes toward each subject were examined in this study using a LISREL correlation model. Reconfirmation of the empirical model was conducted longitudinally based on national data collected from the Longitudinal Study of American Youth (LSAY). The results revealed significant relations of student attitudes among the four school subjects. The strongest correlation was consistently found between student attitudes toward English and social studies. Only after the students reach the 11th grade does the relationship between science and mathematics move into the second highest spot. The empirical attitude relations may be employed by school administrators to coordinate academic programs among different departments, and the indicators of student attitudes can be used by teachers to develop appropriate guidelines to enhance student positive attitudes toward the four major school subjects.

Science, mathematics, English and social studies are the four major subjects in American secondary schools. Students' attitudes toward these subjects are important components of their general attitudes toward secondary education in the U.S. So far, improvement of student attitudes in secondary education has been studied by educators in each of the four subject areas separately (Stephens, 1993). While the four subjects remain separate in secondary schools, research results could be used to guide a joint effort among educators for an integrated improvement of student attitudes toward science, mathematics, English, and social studies. In addition, Grossman and Stodolsky (1995) pointed out: "Shared beliefs about the possibilities and constraints offered by different school subjects help contribute to the 'grammar of schooling' in high schools (Track & Tobin, 1994) and complicate efforts to restructure schools or redesign curriculum" (p. 5).

Nonetheless, little research effort has been expended to date on the examination of student attitude relations toward the study of curricular subjects (Montague, 1993; Siskin, 1991). Thus, an empirical study is needed to assess the strength of the attitude interrelationships based on information collected from a national student sample. The results of such a study might be employed to enhance student positive attitudes toward multiple school disciplines and thus reduce the school dropout rate across the nation. Accordingly, this study was designed to investigate the relations of student attitudes among the four subjects using an empirical data base from an NSF funded project (MDR-8550085), the Longitudinal Study of American Youth (LSAY).

Related Literature

The separation of school courses into subject departments has been the norm in secondary education. Siskin (1991) reported:

Jianjun Wang is Assistant Professor of statistics and educational research, School of Education, California State University, Bakersfield. J. Steve Oliver is Associate Professor of science education, College of Education, The University of Georgia. Andrew T. Lumpe is Assistant Professor of science education, College of Education, University of Toledo. The authors may be contacted at the School of Education, California State University, Bakersfield, 9001 Stockdale Highway, Bakersfield, CA 93311. *This research is supported in part by the University Research Council at California State University, Bakersfield. The data for this study was collected by the Public Opinion Laboratory at Northern Illinois University and was funded by the National Science Foundation (NSF), project MDR-8550085. The views expressed here are not necessarily those of the University Research Council, the Public Opinion Laboratory or the NSF.

> The organization of high schools into departments is a nearly universal feature of the 22,000 secondary schools across the United States; in schools of every location, size, mission, and governance, highly standardized departmental labels divide teachers and courses along academic lines. (p. 134)

Siskin further noted: "Although the academic department is a marketedly familiar feature of the high school, it is also a remarkably unstudied one" (p. 34).

The isolation of school subjects could affect teachers' evaluation about the importance of different school subjects. Grossman and Stodolsky (1995) cautioned: "Pooled analyses of high school teachers, for example, mask the real differences that may exist among teachers of different subjects" (p. 8). They contended: "The issues and concerns of the typical math teacher are not the same as those of the typical English or social studies teacher, nor do they work under the same constraints" (p. 8). Consequently, without proper discretion, teachers' enthusiasm on teaching one subject may convey an egocentric message to students to devalue the importance of other courses.

On the other hand, students are organized by grades across subjects (Siskin, 1991). The information from different instructors could influence students' attitudes toward each subject. Given the condition that teachers value their subjects highly, their inadvertent depreciation of other subjects may lead to conflicting teacher input which hinders integrated enhancement of student positive attitudes. Because of the department separation in secondary schools, the relation of students' attitudes toward different subjects remains as an area of research remarkably unstudied (Siskin, 1991). Further investigation of attitude relations may produce results to prevent the potential attrition of students' positive attitudes caused by the department egocentrism.

Another problem which has impeded the study of attitudes is related to the educational measurement. Many researchers have asserted that it is the method of attitude measurement and the analysis of attitude relations that have lagged behind the present educational practice (Greenwald, 1989). Ostruom (1989), for instance, pointed out:

> Method has affected the study of attitudes in at least two general ways. The very definition of attitude has shifted over the last 40 years, due in large measure to wide-spread adoption of particular methodologies. Second, the techniques themselves have prompted researchers to explore new empirical phenomena. (p.16)

Empirical methods were also supported by many well-known psychologists, including Dawes (1972), Guilford (1954), Likert (1932), and Thurstone (1928, 1931). Dawes (1972) wrote:

> The argument that attitudes cannot be measured because of their intrinsic characteristic likewise rests on a misconception. If an empirical relational system exists, and if an investigator is clever enough to discover or invent a numerical representation of this system, then measurement has in fact occurred. Or if the investigator is able to develop an index measurement technique that leads to predictions that are found to be correct, then index measurement has occurred. The question of whether the investigator is capable of doing either of these things is a purely empirical question. (p.148)

Hence, an empirical approach is adopted in this study to assess the attitude relations of students in American secondary schools.

Moreover, research reports from educators indicated a shift in emphasis among school subjects. Bryant (1982) observed: "With enrollments in the social sciences declining markedly and their educational values being deemphasized, a new approach must be taken to social sciences instruction" (p.1). McConeghy (1987) concurred: "As the technological revolution continues, there is an increased emphasis on mathematics and science" (p.1). As a result, many researchers noticed dramatic changes in student attitude toward secondary schooling (Bryant, 1982; Cain-Caston, 1986; Thorndike-Christ, 1991; White, 1989). To analyze the trend of attitude transition, a longitudinal data base is needed to support the empirical investigation of attitude relations.

The National Center for Education Statistics (NCES) has been collecting longitudinal data across the nation since 1972. Thus far, three projects, National Longitudinal Study of the High School Class of 1972 (NLS-72), High School and Beyond (HS&B), and National Education Longitudinal Study of 1988 (NELS:88), were conducted, but none of them have the extensive coverage of secondary education like the Longitudinal Study of American Youth (LSAY). The school coverage was initially limited to the 12th grade in NLS-72, and subsequently extended to the 10th and 12th grades in HS&B, and the 8th, 10th, and 12th grades in the ongoing NELS:88 project (Davis & Sonnenberg, 1995). Hoffer (1988) reviewed the existing longitudinal studies, and concluded: "Two of the most promising projects currently afield are the National Education Longitudinal Study of

1988 (NELS:88) and the Longitudinal Study of American Youth (LSAY)" (p. 1).

Beginning in the Fall of 1987, the LSAY had national probability samples of approximately 3,000 10th-grade (Cohort 1) and 3,000 7th-grade (Cohort 2) public school students that completed science and mathematics achievement tests and attitudinal questionnaires each year until high school graduation (Miller, 1995). In addition to the comprehensive coverage of secondary education, the LSAY project was built on reliable measurement in education. Hoffer (1988) delineated:

> The NELS88 cognitive tests, for example, included only about half the number of items in the LSAY. And the LSAY attitudinal batteries included at least two and usually three items for each dimension, while the NELS88 batteries have only one item for each dimension. In general, then, the LSAY should measure these dimensions with greater reliability, and the measure should prove more useful for analyses of change over time. (pp. 11-12)

Based on the review of research literature and the existing data sets, the LSAY data base was chosen in this empirical study to assess student attitude relations toward mathematics, science, English, and social studies.

Research Questions

Joreskog and Sorbom (1989) pointed out:

> Most theories and models in the social and behavioral sciences are formulated in terms of theoretical concepts or constructs, which are not directly measurable. However, often a number of indicators or symptoms of such concepts can be used to study the theoretical variables, more or less well. (p. 2)

Although attitude as a construct cannot be measured directly (Henerson, Morris, & Fitz-Gibbon, 1978), students' feedback about subject matter, usefulness, challenge, teacher clarity, difficulty, and textbook clarity can be treated as a set of indicators of student attitude toward a school subject (Wang, Oliver, & Lumpe, 1993). The questions that guide this research are:

1. What relations of student attitudes exist among the four subjects?
2. What longitudinal trends of the attitude relations exist in each student cohort?
3. What are the results of 10th grade comparison based on the student information from Cohort 1 in the first year (Fall, 1987 - Spring, 1988) and Cohort 2 in the last year (Fall, 1990 - Spring, 1991)?

Methods

The LSAY data set contains more than 8,000 variables. Among the survey items are a set of opinionnaire subscales assessing student attitudes toward mathematics, science, English, and social studies. After independently reading the LSAY codebook, and subsequent discussions about relevant indicators of student attitudes, the authors reached agreement regarding which set of variables should be employed to demonstrate student attitudes toward the four school subjects (Table 1). Deletion of outliers and missing values was conducted for these selected variables following the instructions in the LSAY codebook (Miller, Hoffer, Suchner, Brown, & Nelson, 1992).

Table 1
Indicators of Student Attitudes Toward
Math, Science, English and Social Studies

Variable Name	Instrument
Subject Matter	How much do you like the subject matter of each course? A means you really like the subject; F means you hate it.
Teacher Clarity	How clear is the teacher in explaining the material? A means very clear; F means not clear at all.
Challenge	How much does each course challenge you to use your mind? A means that it challenges you a lot; F means that it never challenges you.
Usefulness	How useful do you think each course will be to you in your career? A means that it will be very useful; F means that it will be of no use.
Textbook Clarity	How clear is the textbook for each course? A means very clear; F means not clear at all.
Difficulty	How difficult or easy is each course for you? A means that it is very easy; F means that it is very difficult.

Henerson, Morris, and Fitz-Gibbon (1978) pointed out: "We have no guarantee that the attitude we want to assess will 'stand still' long enough for a one-time measurement to be reliable. A volatile or fluctuating attitude cannot be revealed by information gathered on one occasion" (p. 13). The longitudinal information contained in the LSAY data base was employed to conduct repeated examinations of the relations among student attitudes toward science, mathematics, English and social studies (Table 2).

Table 2
Time Schedule of Repeated Measures of
Student Attitudes in LSAY

	Cohort 1		Cohort 2	
Academic Year	Grade	Semester	Grade	Semester
1987-1988	10	Fall, Spring	7	Fall, Spring
1988-1989	11	Fall, Spring	8	Spring
1989-1990	12	Fall, Spring	9	Fall, Spring
1990-1991			10	Fall, Spring

Variables related to student attitudes are first examined in terms of multicolinearity to eliminate redundant indicators which have been completely covered by the rest of the variables. Joreskog and Sorbom (1984) suggested that the determinant of a correlation matrix is a measure of multicolinearity. They wrote:

> If the determinant is very small relative to the magnitude of the diagonal elements, this is an indication that there are one or more nearly perfect linear relationships among the observed variables. In such a case it is best to delete one or more variables or to use the ULS method instead of the ML method. (p. III.8)

Based on the user's guide of LISREL VI, the determinant of a correlation matrix is desired to be no less than .01 (Joreskog & Sorbom, 1984). Inspection of Table 1 suggested that two indicators, "challenge" and "difficulty" of a subject, are largely overlapped. The first indicator, "subject matter," is ambiguous because the contents of these subjects cover a large body of knowledge base. Student attitudes toward these contents are explicitly exhibited by the tangible responses to the other four items, subject difficulty, teacher clarity, usefulness, and textbook clarity. After deletion of "challenge" and "subject matter" items, determinants of the correlation matrices are all larger than .01 (Table 3). Thus, the four remaining variables, "difficulty," "teacher clarity," "usefulness," and "textbook clarity" are employed as indicators of student attitudes toward the four subjects.

Table 3
Determinants of Correlation Matrices
Among the Attitude Indicators

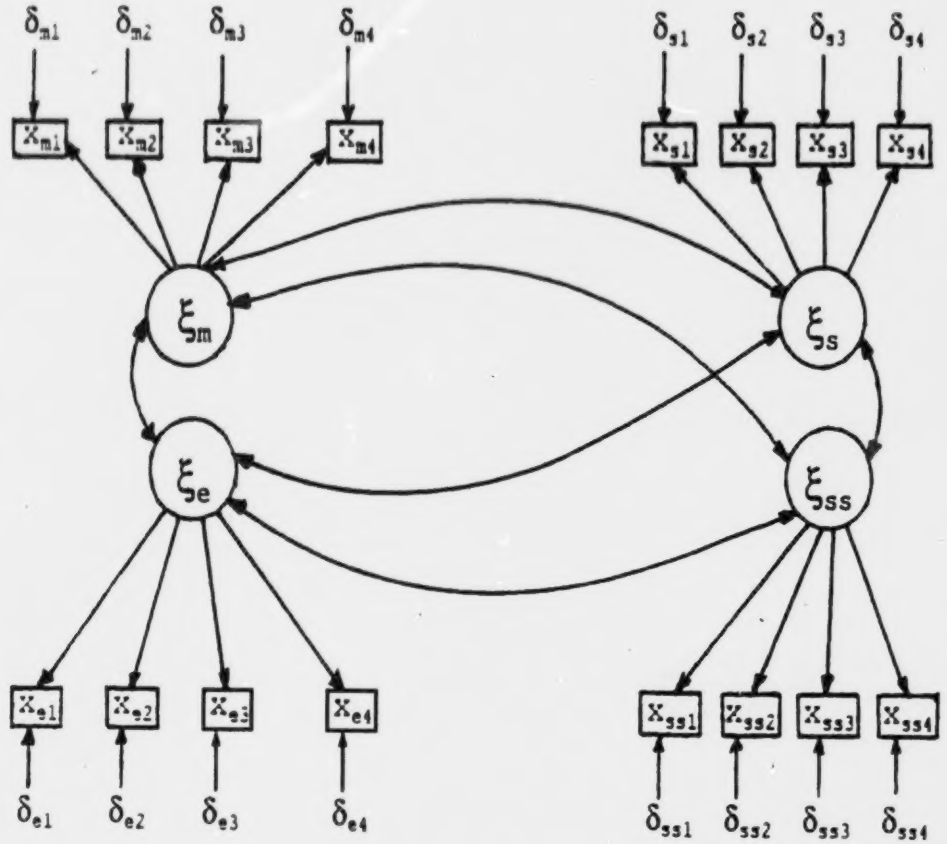
Measurement	Cohort 1	Cohort 2
1987 Fall	0.0822767	0.0689836
1988 Spring	0.0704761	0.0574902
1988 Fall	0.0491228	
1989 Spring	0.0518645	0.0610162
1989 Fall	0.0703813	0.0565641
1990 Spring	0.0476912	0.0459217
1990 Fall		0.0566272
1991 Spring		0.0483775

Student attitudes toward each subject were treated as latent variables, and empirically identified through factor analysis (Joreskog & Sorbom, 1993). Joreskog and Sorbom (1993) presumed: "Latent variables are unobservable and have no definite scale" (p. 7). Thus, not all central tendency statistics, such as means and variances, are estimable. According to Joreskog and Sorbom (1993), "The most useful and convenient way of assigning the units of measurement of the latent variables is to assume that they are standardized so that they have unit variances in the population" (p. 7). Under this LISREL assumption, the focus of this study is on the attitude relations, rather than the central tendency statistics. To accom-modate all potential relations of student attitudes, no pre-conditions are imposed to restrict the correlations of attitudes among the four subjects. Thus, six or C_4^2 pairwise correlation coefficients are estimated through the maximum likelihood method, and the corresponding t-values are calculated to examine the existence of these correlation coefficients (Joreskog & Sorbom, 1989). The whole model of investigation is summarized in a path diagram (Figure 1).

Goodness of Fit Index (GFI) and Root-Mean-Square Residual (RMR) were adopted to evaluate the model fitness with the empirical data. The total coefficient of determination is computed to measure how well the indicators jointly serve as measurement instruments for the four attitude constructs (Figure 1: ξ_m, ξ_e, ξ_s, and ξ_{ss}). These results are examined over the two cohorts to facilitate the identification of longitudinal trends in each cohort and the 10th grade comparison between the two cohorts.

Results

The correlation coefficients are tabulated for the four subjects over four years (Table 4), and the corresponding t-values are listed in Table 5.


δ_{m1}
δ_{m2}
δ_{m3}
δ_{m4}
X_{m1}
X_{m2}
X_{m3}
X_{m4}
δ_{s1}
δ_{s2}
δ_{s3}
δ_{s4}
X_{s1}
X_{s2}
X_{s3}
X_{s4}
ξ_m
ξ_s
ξ_e
ξ_{ss}
X_{e1}
X_{e2}
X_{e3}
X_{e4}
X_{ss1}
X_{ss2}
X_{ss3}
X_{ss4}
δ_{e1}
δ_{e2}
δ_{e3}
δ_{e4}
δ_{ss1}
δ_{ss2}
δ_{ss3}
δ_{ss4}

Mathematics
X_{m1}: teacher clarity
X_{m2}: usefulness
X_{m3}: textbook clarity
X_{m4}: difficulty
ξ_m : attitude toward math

English
X_{e1}: teacher clarity
X_{e2}: usefulness
X_{e3}: textbook clarity
X_{e4}: difficulty
ξ_e : attitude toward English

Science
X_{s1}: teacher clarity
X_{s2}: usefulness
X_{s3}: textbook clarity
X_{s4}: difficulty
ξ_s : attitude toward science

Social Studies
X_{ss1}: teacher clarity
X_{ss2}: usefulness
X_{ss3}: textbook clarity
X_{ss4}: difficulty
ξ_{ss} : attitude toward social studies

$\delta_{m1}, \delta_{m2}, \ldots, \delta_{ss4}$: measurement errors of the corresponding indicators

Figure 1. Path Diagram of the Attitude Model

BEST COPY AVAILABLE

Table 4
Correlation Coefficients of Student Attitudes Toward Mathematics, Science, English, and Social Studies*

	Correlation Coefficients					
Measurement	Eng-Soc	Math-Sci	Math-Eng	Math-Soc	Sci-Eng	Sci-Soc
Cohort 1						
1987 Fall	0.420	0.237	0.235	0.231	0.363	0.260
1988 Spring	0.498	0.264	0.322	0.315	0.397	0.289
1988 Fall	0.522	0.374	0.261	0.236	0.344	0.371
1989 Spring	0.583	0.419	0.329	0.317	0.338	0.411
1989 Fall	0.587	0.433	0.395	0.231	0.428	0.329
1990 Spring	0.684	0.468	0.368	0.268	0.377	0.383
Mean	0.555	0.369	0.319	0.266	0.375	0.332
Median	0.553	0.397	0.326	0.252	0.370	0.328
Cohort 2						
1987 Fall	0.604	0.429	0.463	0.400	0.494	0.527
1988 Spring	0.568	0.392	0.469	0.432	0.505	0.466
1989 Spring	0.621	0.536	0.547	0.495	0.508	0.557
1989 Fall	0.664	0.443	0.437	0.363	0.479	0.440
1990 Spring	0.540	0.506	0.494	0.343	0.393	0.493
1990 Fall	0.569	0.358	0.480	0.393	0.478	0.498
1991 Spring	0.417	0.233	0.219	0.225	0.349	0.260
Mean	0.573	0.418	0.449	0.382	0.460	0.467
Median	0.569	0.429	0.469	0.393	0.479	0.493

* Fisher's z transformation (Johnson & Wichern, 1988) was used when computing means and medians of the correlation coefficients

Table 5
t-Values for the Correlation Coefficients among the Student Attitudes

	t-Values					
Measurement	Eng-Soc	Math-Sci	Math-Eng	Math-Soc	Sci-Eng	Sci-Soc
Cohort 1						
1987 Fall	15.227	8.386	8.157	8.166	12.813	9.050
1988 Spring	14.818	7.602	9.451	8.677	12.005	7.863
1988 Fall	16.632	11.350	7.654	6.622	10.312	10.818
1989 Spring	18.945	12.825	9.746	9.154	9.790	11.995
1989 Fall	16.319	10.187	9.900	5.314	10.394	7.466
1990 Spring	30.148	16.437	13.505	9.196	12.996	12.795
Cohort 2						
1987 Fall	20.074	13.229	13.715	11.749	15.791	17.770
1988 Spring	19.030	12.497	15.081	13.804	16.564	15.153
1989 Spring	20.647	16.875	16.977	14.490	16.565	18.189
1989 Fall	22.349	13.196	12.610	10.376	14.400	13.297
1990 Spring	18.631	16.341	15.987	10.238	12.315	16.315
1990 Fall	17.740	10.258	15.027	11.152	14.528	14.520
1991 Spring	23.499	11.257	11.955	9.180	14.313	18.282

To verify the model fitness, Goodness of Fit Index (GFI) and Root-Mean-Square Residual (RMR) are assembled in Tables 6 and 7 respectively. The total coefficients of determination based on the four LSAY indicators in each subject are listed in Table 8.

Table 6
Goodness of Fit Index (GFI) of the Attitude Model

Measurement	Cohort 1	Cohort 2
1987 Fall	0.928	0.929
1988 Spring	0.922	0.914
1988 Fall	0.894	
1989 Spring	0.900	0.918
1989 Fall	0.902	0.894
1990 Spring	0.889	0.915
1990 Fall		0.913
1991 Spring		0.909

Table 7
Root-Mean-Square Residual (RMR) of the Attitude Model

Measurement	Cohort 1	Cohort 2
1987 Fall	0.058	0.055
1988 Spring	0.057	0.060
1988 Fall	0.067	
1989 Spring	0.066	0.058
1989 Fall	0.065	0.067
1990 Spring	0.067	0.060
1990 Fall		0.061
1991 Spring		0.063

Table 8
Total Coefficient of Determination of the Attitude Indicators

Measurement	Cohort 1	Cohort 2
1987 Fall	0.968	0.951
1988 Spring	0.968	0.958
1988 Fall	0.964	
1989 Spring	0.960	0.952
1989 Fall	0.955	0.943
1990 Spring	0.959	0.971
1990 Fall		0.958
1991 Spring		0.966

Discussion

Byrne (1989) pointed out:

> One of the initial things to look at when searching for misfit in a model is to examine the statistical significance of each parameter. Nonsignificant parameters can be considered unimportant to the model and can be subsequently fixed to a value of 0.0; they are thereby deleted from the model. The statistical significance of parameters can be determined by examining the t-values provided by LISREL. These values represent the parameter estimate divided by its standard error. As such, t-values provide evidence of whether or not a parameter is significantly different from zero; values >2.00 are generally considered to be statistically significant. (p.56)

T-values listed in Table 5 are much larger than 2.00. Accordingly, all six positive correlation coefficients are significant regardless of the semester and cohort of the measurement. Inspection of Table 4 also suggests that the correlation coefficients ranged from .219 to .684 and are too large to be interpreted by the effect of randomization. Hence, positive relations of student attitudes have been found among the four major subjects in U.S. secondary schools based on the LSAY data analysis.

An examination of the longitudinal trend in Table 5 uncovers a consistent pattern that the strongest correlation of student attitudes exists between English and social studies. In contrast, the correlations between mathematics and science were not consistently ranked the second largest until the students reached the 11th grade (Table 4: Cohort 1, 1988 Fall). In addition, the 10th grade information was collected twice in LSAY from Cohort 1 in the first year (Fall, 1987 - Spring, 1988) and Cohort 2 in the last year (Fall, 1990 - Spring, 1991). The results from both cohorts revealed the second highest correlations between science and social subjects, such as English and social studies. The consistency of the longitudinal pattern seemed to suggest that a joint effort from educators in different subjects could have a proliferating impact on the improvement of student attitudes toward science, mathematics, English and social studies.

The Goodness of Fit Index (GFI) designates the relative amount of variance and covariance jointly explained by the model, and the Root-Mean-Square Residual (RMR) denotes the average discrepancy between the

BEST COPY AVAILABLE

elements in the sample and hypothesized covariance matrices (Byrne, 1989). The high GFI in Table 6 and low RMR in Table 7 congruously verified a fairly good fit between the attitude model (Figure 1) and the LSAY data base. The remarkably high coefficients of determination (Table 8) also provided a strong indication that the four LSAY indicators jointly served as good instruments for measuring the latent student attitudes.

In summary, a high-quality data base was chosen in this empirical study after an extensive review of the existing longitudinal data sets released by the NCES and the NSF. The identification of the attitude constructs and the estimation of the correlation coefficients were conducted using a well-developed LISREL computer software program (Joreskog & Sorbom, 1984, 1989, 1993). The empirical results consistently reconfirmed the significant relations of student attitudes among the major school subjects. Among the findings of this empirical study, the strength of the relations may be employed by school administrators to coordinate academic programs among different departments, and the indicators of student attitudes (Table 1) can be used by teachers to develop appropriate guidelines, such as the following, to enhance student positive attitudes:

1. Express interest in all major school subjects;
2. Encourage students to meet the challenge of each course;
3. Clarify textbook confusion through well-prepared lectures, and whenever pertinent, present the subject content in an interdisciplinary context;
4. Stress confidence in students' abilities in mathematics, science, English, and social studies;
5. Appreciate the usefulness of each subject.

References

Byrne, B. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models.* NY: Springer-Verlag.

Bryant, H. A. (1982, March). *The social sciences: Towards a new approach to teaching them.* Paper presented at the annual conference of the Community College Social Science Association, Las Vegas, NV.

Cain-Caston, M. (1986). *Parent and student attitudes toward mathematics as they relate to third grade mathematics achievement.* North Carolina, U.S.: ERIC Clearing House. (ERIC Document Reproduction Service, No. ED334078).

Davis, C., & Sonnenberg, B. (1995). *Programs and plans of the National Center for Education Statistics.* Washington, DC: U.S. Department of Education.

Dawes, R. M. (1972). *Fundamentals of attitude measurement.* NY: John Wiley & Sons.

Greenwald, A. G. (1989). Why are attitudes important? In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Grossman, P. L., & Stodolsky, S. S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(8), 5-11, 23.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). NY: McGraw-Hill.

Henerson, M., Morris, L., & Fitz-Gibbon, C. (1978). *How to measure attitudes.* London: Sage Publications.

Hoffer, T. B. (1988, April). *The Longitudinal Study of American Youth and the National Education Longitudinal Study of 1988: A comparison.* Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis,* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Joreskog, K. G., & Sorbom, D. (1984). *LISREL VI user's guide.* Mooresville, Indiana: Scientific Software.

Joreskog, K. G., & Sorbom, D. (1989). *LISREL 7 a guide to the program and applications,* (2nd ed.). Chicago, IL: SPSS.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 8 New.* Chicago, IL: Scientific Software International.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44-53.

McConeghy, J. L. (1987, April). *Mathematics attitudes and achievement: Gender differences in a multivariate context.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Miller, J. (1995). *Longitudinal Study of American Youth -- Overview of study design and data resources.* Chicago, IL: The Chicago Academy of Sciences.

Miller, J. D., Hoffer, T., Suchner, R., Brown, K. G., & Nelson, C. (1992). *LSAY codebook.* DeKalb, IL: Northern Illinois University.

Montague, M. (1993). Student-centered or strategy-centered instruction: What is our purpose? *Journal of Learning Disabilities, 26*(7), 433-437, 481.

Ostruom, T. M. (1989). Interdependence of attitude theory and measurement. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Siskin, L. S. (1991). Departments as different worlds: Subject subcultures in secondary schools. *Educational Administration Quarterly, 27*(2), 134-160.

Stephens, G. (1993). Wrong questions lead to misdirected answers. *School Administrator, 50*, 10-11.

Thorndike-Christ, T. (1991). *Attitudes toward mathematics: Relationships to mathematics achievement, gender, mathematics course-taking plans, and career interests.* Washington, U.S.: ERIC Clearing House. (ERIC Document Reproduction Service, No. ERIC: ED347066).

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Thurstone, L. L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26*, 249-269.

Track, D. & Tobin, W. (1994). The "grammar" of schooling: Why has it been so hard to change? *American Educational Research Journal, 31*, 453-479.

Wang, J., Oliver, J. S., & Lumpe, A. T. (1993, April). *A longitudinal study of American youth: Students' attitudes toward science, math, English and social studies.* Paper presented at the annual meeting of the National Association for Research in Science Teaching (NARST), Atlanta, GA.

White, N. (1989). *Developmental relationships between students' attitudes toward reading and reading achievement in grades 1 through 8.* Iowa, U.S.: University of Northern Iowa. (ERIC Document Reproduction Service, No. ERIC: ED329905).

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 23-33

The Verbal and Nonverbal Intelligence of American vs. French Children at Ages 6 to 16 Years

Alan S. Kaufman
Psychological Assessment Resources, Inc. (PAR)

James C. Kaufman
Yale University

Nadeen L. Kaufman
California School of Professional Psychology, San Diego

Mireille Simon
Les Editions du Centre de Psychologie Appliqué (ECPA), Paris

The standardization samples (ages 6 to 16 years) for the American WISC-R (N = 2,200) and the French WISC-R (N = 1,066) were used as the data base for comparing the intelligence of children living in the United States with the intelligence of children living in France. Four Verbal subtests and six Performance subtests included the same, or virtually the same, items in both versions of the WISC-R. Normative data presented in the respective test manuals provided the necessary information for comparing the intelligence test performance of the two cultures. Average scores earned by French children on the 10 subtests were entered into the American norms to permit comparisons between French and American children. Verbal IQs were prorated based on the four subtests, and Performance IQs were computed using the five regular subtests. Because the standardizations occurred about 6-1/2 years apart, mean IQs and scaled scores were adjusted for generational changes, as reported by Flynn. The total sample of American children scored significantly higher than the total sample of French children on Verbal IQ, Full Scale IQ, and the following subtests: Similarities, Comprehension, Digit Span, Picture Completion, Picture Arrangement, and Coding. French children scored significantly higher on Block Design. No significant differences were observed on Performance IQ, Arithmetic, Object Assembly, or Mazes. Implications of these results were discussed.

Cross-cultural comparisons of intelligence are often difficult to conduct because of differences in instrumentation and sampling, although the results of these studies frequently arouse the interest of laypersons as well as scientists. Some research results have proved to be controversial and newsworthy, such as Lynn's (1982) initial study that reported a mean WISC-R IQ of 111 for Japanese children and adolescents when compared to American norms. Methodological problems pointed out by Flynn (1983), Stevenson and Azuma (1983), and Vining (1983), caused Lynn (1983) to rethink his position and, ultimately, to reanalyze the data (Lynn & Hampson, 1986c). The latter analysis suggested that Japanese children scored 2½ points higher than American children, not 11 points as originally reported, although only the initial finding aroused the attention of the media.

The pitfalls in Lynn's (1982) first study, delineated by researchers from a variety of vantage points, helped provide guidelines for subsequent cross-cultural research. Vining (1983) pointed to the need to consider differential variability of scores within cultures, and Stevenson and Azuma (1983) stressed the impact of variables such as urban-rural residence and socioeconomic status when they are used to stratify one sample and not the other. Flynn (1983) emphasized the need to control for generational shifts in IQ when the standardizations are conducted in different years, to utilize data from subtests even if they have undergone minor modifications, and to be sure to select the appropriate mean score for comparison. In the study of a homogeneous population

---

Alan S. Kaufman is a Senior Research Scientist for Psychological Assessment Resources, Inc. (PAR); James C. Kaufman is a Teaching Fellow in Psychology at Yale University; Nadeen L. Kaufman is a Professor of Clinical Psychology at the California School of Professional Psychology, San Diego; and Mireille Simon is Deputy Managing Director at Les Editions du Centre de Psychologie Appliqué (ECPA) in Paris, France. Address correspondence regarding the article to James C. Kaufman, Teaching Fellow, PO Box 208205, 2 Hillhouse Avenue, Yale University, New Haven, CT 06520-8205.

such as the Japanese, Flynn argued (and Lynn concurred), the appropriate mean for comparison is the mean of about 102 for white children, not the overall mean of 100.

Cross-cultural comparisons in intelligence between the United States and other countries have more frequently involved Asian countries such as Japan, China, and Korea than European countries (Ishikuma, 1990; Kaufman, McLean, Ishikuma, & Moon, 1989; Lynn & Hampson, 1986b, 1986c; Moon, 1988; Stevenson et al., 1985). Results of previous studies have suggested relatively small differences in overall cognitive ability between Americans and Asians, although differences in *patterns* have been noted. Japanese children seem to excel in simultaneous, visual-spatial processing of information, relative to Americans, but American children tend to perform better on verbal and sequential tasks (Ishikuma, Moon, & Kaufman, 1988; Kaufman et al., 1989; Lynn & Hampson, 1986b, 1986c; Stevenson et al., 1985). In contrast, Korean children have a strength, relative to Americans, in the ability to solve sequential step-by-step problems (Moon, 1988; Moon, Byun, McLean, & Kaufman, 1994), a strength that may be characteristic of Chinese children as well (Stevenson et al., 1985).

The present investigation examined American-French cross-cultural differences at ages 6 to 16 years on the Wechsler Intelligence Scale for Children (WISC-R; Wechsler, 1974). The data base comprised normative data presented in the respective test manuals for two standardization samples, one tested in the United States in the early 1970s during the original standardization of the WISC-R (Wechsler, 1974), and the other tested in France in the late 1970s during the norming of the French translation and adaptation of the WISC-R (Wechsler, 1981). The goals of this study were: (a) to identify all subtests that were either retained intact in the French version of the WISC-R, or modified only slightly, to provide a basis of comparison of the intellectual abilities of children living in the two countries; (b) to compare performance on the WISC-R IQs and pertinent subtests, taking into account the methodological variables pointed out by Flynn (1983) and others that might compromise the validity of the data; and (c) to investigate Verbal-Performance IQ differences, as well as subtest profile differences, to determine whether American and French children display different patterns of intelligence, even if overall scores are comparable.

Method

Subjects

The American standardization of the WISC-R included a total of 2,200 children and adolescents ages 6 to 16 years (Wechsler, 1974). The sample included 100 boys and 100 girls at each year of age between 6 and 16. Each individual tested was within six weeks of his or her half-birthday. The sample was stratified on the variables of age, gender, geographic region, race (white-nonwhite), and socioeconomic status (occupation of head of household). Testing was conducted from December 1971 to January 1973.

The French standardization of the WISC-R included a total of 1,066 children and adolescents ages 6 to 16 years, who were tested from April 1978 to June 1979 (Wechsler, 1981). The sample included 50 boys and 50 girls at each year of age between 6 and 11, and equal numbers of males and females at ages 12 ($N = 98$), 13 ($N = 94$), 14 ($N = 98$), 15 ($N = 92$), and 16 ($N = 84$). As in the U. S. sample, each individual tested was within six weeks of his or her half-birthday. The French sample was stratified on the variables of age, gender, community size, socioeconomic status (father's occupation), and grade in school. For the last variable, individuals were selected for the sample to try to ensure that proportional numbers of each age group were in the grade that was appropriate for their age, in a grade intended for children younger than their age, and in a grade intended for children older than their age.

In France, the person's language (French or non-French), rather than race, is of primary importance. Race, therefore, was considered irrelevant as a stratification variable by the test publishers; in any case, data on race are unavailable in France. Data are also unavailable on the percentage of children whose home language is not French, although data are provided for the percentage of French versus non-French (i.e., foreign) individuals. According to the French Department of Education, approximately 10% of children between the ages of 6 and 18 years are not French. The groups who are not strictly of French origin include Portuguese, Algerians, Tunisians, Spanish, Africans, Moroccans, Cambodians, Vietnamese, and others. The heterogeneity of the French sample on this variable is generally comparable to the heterogeneity of the American sample on the variable of race (15% nonwhite).

Instruments

The WISC-R subtests that were barely changed or unchanged from the American to the French version were used to examine cross-cultural differences. The following subtests were retained intact in the French version of the WISC-R, including the scoring systems (e.g., application of bonus points for quick, perfect performance): Picture Arrangement, Block Design, Object Assembly, Coding, and Mazes. Digit Span utilized the identical items, although the numbers were read in French instead of English and, therefore, presented a different set of linguistic (as opposed to numeric) stimuli to be repeated. Picture Completion included one different item, a commode missing its "bouton" (handle) instead of a ladder missing its step. Similarities included direct translations of all items except for the necessary substitution of METRE-KILO for POUND-YARD. Arithmetic included the same mathematical operations for all but one item. The next-to-hardest item (one of the three that have the words printed on a card) was changed completely to an item of apparently comparable difficulty. Otherwise, translations of the word problems were, in nearly all items, precise, except for the use of "francs" and French names. Comprehension included direct translations of 15 of the 17 items. Two items were changed. French items regarding the advantages of "cassettes" over "disques" and the reasons for having "publicité" replaced American items about paperbacks versus hard-cover books and the use of cotton for making cloth.

The above-named 10 WISC-R subtests were judged by the authors of this article to be sufficiently similar in the two versions of the WISC-R to permit cross-cultural comparisons. The modifications were considered slight and not likely to have a significant impact on cross-cultural comparisons. However, Information and Vocabulary were excluded from the study because item changes in each subtest were numerous.

Procedure

The average raw score obtained by French children for each of the 11 age groups on each of the 10 subtests in the analysis was estimated by analyzing the raw scores that correspond to a scaled score of 10 in the French norms tables. When a single raw score corresponded to a scaled score of 10, then that raw score was considered the average raw score. When more than one raw score corresponded to a scaled score of 10, then the midpoint of the raw scores was considered the average. When no raw score corresponded to 10, then the mean of the raw scores corresponding to scaled scores of 9 and 11 was considered the average.

These averages were determined for the French norms tables corresponding to each half-year (e.g., 6 years, 4 months, 0 days to 6 years, 7 months, 30 days) because those norms groups provided the closest match to the actual ages of the children in the two standardization samples.

The average scores earned by the French children on each subtest at each age were then entered into the appropriate American norms table, and the corresponding average scaled score was obtained by interpolation. Mean French scaled scores, relative to American norms, were computed for each of the 10 subtests, by age, and for the total sample. Mean Verbal sums of scaled scores were obtained by prorating the sums of the mean scores on the four Verbal subtests, and the prorated Verbal sums were entered into the American IQ conversion table to obtain mean Verbal IQs. Mean Performance sums of scaled scores were computed by summing the mean scaled scores on the five regular subtests (excluding Mazes), and these sums were entered into the American IQ conversion table to obtain mean Performance IQs. The Verbal and Performance sums of scaled scores were added together, and those Full Scale sums were entered into the American IQ conversion table to obtain mean Full Scale IQs.

All of these mean scores had to be adjusted for changes in the American norms due to generational changes in the intelligence of American children occurring between the standardization of the American WISC-R in 1971-73 and the standardization of the French WISC-R in 1978-79 (Flynn, 1984, 1987). The midpoint of the American standardization was June/July 1972, an average of 6 years 4.5 months earlier than the French standardization (midpoint = November, 1978). Based on Flynn's (1987) data for Americans on the WISC and WISC-R, converted to a common metric by Kaufman (1990, Table 2.4), American IQs increase at a rate of 3.0 points per decade (2.7 Verbal and 3.3 Performance). These values were used to adjust the means earned by the French based on American norms, multiplying the values by .6375 (the interval between the two standardizations, i.e., 6.375 years). Hence, 1.72 points (.115 *SD*) was subtracted from each French Verbal IQ, 2.10 points (.140 SD) was subtracted from each French Performance IQ, and 1.91 points (.127 *SD*) was subtracted from each French Full Scale IQ. When adjusting scaled scores by the same relative amount, all Verbal scaled scores were adjusted by the amount of change in Verbal IQ (.115 *SD*) and all Performance scaled scores were adjusted by the amount of change in Performance IQ (.140 *SD*). Therefore, .34 point was subtracted from each French Verbal scaled score, and .42 point was subtracted from each French Performance scaled score.

It is important to note that even though Flynn's (1984, 1987) data for children were based on data obtained from the 1930s to the early 1970s, subsequent research with children has indicated that the same generational change of about 3 IQ points per decade has persisted as an apparent built-in constant in the U. S.

(Kaufman & Kaufman, 1983, Table 4.19; Wechsler, 1991, Table 6.8). It was sensible, therefore, to apply Flynn's corrections to the data analyzed in the present study which were gathered both in the early and late 1970s.

After obtaining mean scores for French children relative to American norms, the next step was to determine whether French and American children differed significantly in their mean scores. Standard deviations for the French sample were set equal to the standard deviations reported for the French sample in the manual for the French WISC-R (Wechsler, 1981, Tables 4 and 12). Subtest *SD*s for the total sample were computed by taking the square root of the mean variance for the 11 age groups on each subtest. The *SD*s for the IQs were computed from the *SD*s of the sums of scaled scores by setting the values reported for the total sample equal to 15.0 and computing the age-by-age values as a proportion of the *SD* of the total sample.

Means and *SD*s had to be determined for the American sample. These values were obtained from data reported in the American WISC-R manual (Wechsler, 1974, Tables 6 and 14). Subtest and IQ Scale *SD*s were obtained using the methods described for the French sample. Mean scaled scores were obtained from values reported in Table 14 of the WISC-R manual. Mean IQs were obtained by entering the mean sum of scaled scores for the Verbal, Performance, and Full Scales (Wechsler, 1974, Table 14) into the pertinent IQ conversion table for Americans and interpolating.

Vining's (1983) criticism of Lynn's (1982) research, namely that Japanese children produced more homogeneous distributions of scores than did American children, was not true for the French versus American comparison. Mean sums of scaled scores for the three IQ scales, which reflect the variability of the test scores on the global scales within each culture, produced remarkably similar *SD*s for American and French children at ages 6 to 16 years (Wechsler, 1974, Table 6; Wechsler, 1981, Table 4). The *SD*s for the total American and French samples were as follows: Verbal Scale (12.14/12.43), Performance Scale (10.89/10.93), and Full Scale (21.01/20.96).

Design

The 12 sets of mean Verbal IQs (one per age plus total sample) were compared using two-tailed t tests for independent samples. The Bonferroni correction was applied to control for the chance errors that are introduced when making several comparisons simultaneously. To correct for the 12 comparisons, $t = 3.04$ was required for significance to achieve a familywise alpha level of .05, and a $t = 3.35$ was required for a familywise alpha level of .01. The same procedures were followed to compare the 12 sets of mean Performance and Full Scale IQs. For the 10 subtests, only the values for the total samples (i.e., the means for the 11 age groups) were compared statistically. For 10 simultaneous comparisons, a Bonferroni-corrected $t = 2.97$ was required for significance to achieve a familywise alpha level of .05, and $t = 3.29$ was required for a familywise alpha level of .01.

It was arbitrary in this investigation whether to enter French scores into American norms, or vice versa. American norms were used, following procedures used in previous investigations (Lynn, 1982; Lynn & Hampson, 1986b; Moon, 1988). When using American norms, it is appropriate to correct mean differences for generational changes within the United States, because such changes make the norms out of date by a predictable number of points per year. If the French norms are used, however, then the mean differences should be changed in accordance with generational changes within *France*.

Flynn's (1987) research indicates that generational changes in France are more substantial than changes in the U. S., and that changes in France are much larger on nonverbal than verbal measures. Flynn's (1987) data for the French, converted to a common metric by Kaufman (1990, Table 2.4), indicate that IQs increase at a rate of 5.3 points per decade (2.35 verbal and 8.15 nonverbal). These increases reflect the average of data for two studies reported by Flynn, one using Raven's matrices and a test of verbal and math ability, and the other using the French WISC and WISC-R.

Therefore, the results of the present study will be slightly different whether using French versus American norms, primarily because the adjustments to the mean differences will vary as a result of which culture is used to define generational changes. The examination of V-P IQ differences was a specific purpose of this study, yet such discrepancies will be greatly affected if the data are adjusted for American generational differences (V and P change about the same amount) or for French generational changes (P changes much more than V).

Consequently, some analyses were repeated by entering means for American samples into the French norms, and adjusting the differences for generational changes in France. Only the IQs, not the scaled scores, were recomputed, and no tests of statistical significance were run, to limit the number of comparisons made. However, the results were used to provide a context for interpreting the original findings. Based on Flynn's (1987) data for the French, the differential of 6 years 4-1/2 months between the standardizations of the American and French WISC-Rs was controlled by adding 1.50 points to each American Verbal IQ, 5.20 points to each

American Performance IQ, and 3.38 points to each American Full Scale IQ.

Results

Table 1 shows the mean WISC-R IQs for French children, relative to American norms, adjusted for generational changes within the United States. American children scored significantly higher ($p < .01$) for all ages combined and for five of the 11 age groups. Differences ranged from about 4 to 7 points, with no age-related changes evident. The overall difference of about 5 points (.35 *SD*) represents the 11 age groups quite well; separate differences for the 11 subsamples should not be interpreted since several values either just missed, or just qualified, for statistical significance.

Performance IQ and Full Scale IQ differences are presented in Tables 2 and 3, respectively. None of the differences for Performance IQ reached significance at the .05 level, and only the Full Scale IQ difference of about 2 points (.16 *SD*) for the total sample, favoring the Americans, achieved significance. The global results indicate higher Verbal IQ for American than French children across the age range and no meaningful cross-cultural difference in Performance IQ. The significant difference on the Full Scale, therefore, simply reflects the apparent verbal advantage of the American sample, as well as the fact that even small differences can produce statistical significance with large samples; it is not a meaningful finding.

Table 4 presents mean differences on the 10 subtests analyzed for this study, and the findings help explain the results of the IQ analyses. The higher Verbal IQ by American children reflects their better performance on three of the four subtests and denotes a consistent superiority on tests of short-term memory (Digit Span), verbal concept formation (Similarities) and reasoning in social situations (Comprehension). The subtest analysis indicates that the failure of Performance IQ to produce significant cross-cultural differences is not a function of equal ability by American and French children in the diverse aspects of nonverbal intelligence. In fact, the nearly equal IQs mask notable differences within the subtest profile.

American children scored significantly higher on Picture Completion, Picture Arrangement, and Coding, whereas French children scored significantly higher on Block Design, and had higher mean scores on Mazes and Object Assembly that approached significance. The largest subtest discrepancies were on Digit Span (.47 *SD*) and Similarities (.33 *SD*) favoring American children, and on Block Design (.30 *SD*), favoring French children.

Table 1
Comparison of French and American Intelligence on the WISC-R Verbal Scale, Entering Mean French Scores into American Norms--ADJUSTED FOR GENERATIONAL CHANGES IN THE U. S.

	American Verbal IQ			French Verbal IQ			Mean	*t* of
Age	N	Mean	SD	N	Mean	SD	Diff.	Diff.
6.5	200	99.16	13.96	100	93.94	13.67	+5.22	3.07*
7.5	200	99.06	14.49	100	94.16	13.62	+4.90	2.82
8.5	200	100.72	14.01	100	95.40	14.59	+5.32	3.06*
9.5	200	99.68	15.78	100	96.03	14.94	+3.65	1.92
10.5	200	100.32	14.14	100	96.58	14.99	+3.74	2.12
11.5	200	101.23	15.51	100	94.47	14.31	+6.76	3.65**
12.5	200	99.90	15.14	98	95.09	15.33	+4.81	2.57
13.5	200	100.56	15.70	94	94.57	15.13	+5.99	3.09*
14.5	200	100.22	15.62	98	94.88	16.42	+5.34	2.73
15.5	200	100.63	14.43	92	94.78	16.21	+5.85	3.09*
16.5	200	100.17	16.07	84	95.40	16.30	+4.77	2.27
MEAN	2200	100.25	15.00	1066	95.03	15.00	+5.22	9.33**

$^{*}p < .05$ $^{**}p < .01$

Note. Bonferroni-corrected *t* values of 3.04 and 3.35 are required for familywise alpha levels of .05 and .01, respectively.

BEST COPY AVAILABLE

Table 2
Comparison of French and American Intelligence on the WISC-R Performance Scale, Entering Mean French Scores into American Norms--ADJUSTED FOR GENERATIONAL CHANGES IN THE U. S.

	American Performance IQ			French Performance IQ			Mean	*t* of
Age	N	Mean	SD	N	Mean	SD	Diff.	Diff.
6.5	200	99.76	15.99	100	100.09	14.89	-0.33	-0.17
7.5	200	99.62	14.92	100	102.02	15.34	-2.40	-1.30
8.5	200	100.24	14.78	100	101.95	15.44	-1.71	-0.93
9.5	200	100.66	14.86	100	99.98	14.70	+0.68	0.37
10.5	200	99.48	13.84	100	99.99	13.63	-0.51	-0.30
11.5	200	100.67	14.92	100	99.98	14.59	+0.69	0.38
12.5	200	99.18	15.45	98	99.96	14.60	-0.78	-0.42
13.5	200	101.05	15.40	94	99.80	13.90	+1.25	0.67
14.5	200	99.26	14.70	98	99.57	16.44	-0.31	-0.16
15.5	200	100.56	15.01	92	98.26	14.97	+2.30	1.22
16.5	200	100.24	15.14	84	98.54	17.03	+1.70	0.83
MEAN	2200	100.19	15.00	1066	100.01	15.00	+0.18	0.32

Note. Bonferroni-corrected *t* values of 3.04 and 3.35 are required for familywise alpha levels of .05 and .01, respectively.

Table 3
Comparison of French and American Intelligence on the WISC-R Full Scale, Entering Mean French Scores into American Norms--ADJUSTED FOR GENERATIONAL CHANGES IN THE U. S.

	American Full Scale IQ			French Full Scale IQ			Mean	*t* of
Age	N	Mean	SD	N	Mean	SD	Diff.	Diff.
6.5	200	99.31	14.85	100	97.12	13.91	+2.19	1.23
7.5	200	99.23	14.66	100	97.76	14.40	+1.47	0.82
8.5	200	100.23	14.26	100	98.43	15.21	+1.80	1.01
9.5	200	100.00	15.59	100	98.29	14.58	+1.71	0.91
10.5	200	99.71	14.01	100	98.73	14.53	+0.98	0.56
11.5	200	100.70	15.41	100	97.33	13.92	+3.37	1.84
12.5	200	99.38	15.58	98	97.71	14.79	+1.67	0.88
13.5	200	100.56	15.47	94	97.21	14.43	+3.35	1.77
14.5	200	99.57	15.25	98	97.27	16.84	+2.30	1.18
15.5	200	100.34	14.41	92	95.95	15.79	+4.39	2.35
16.5	200	99.94	15.66	84	97.85	17.12	+2.09	1.00
MEAN	2200	99.96	15.00	1066	97.60	15.00	+2.36	4.22**

**$p < .01$

Note. Diff. = Difference. Bonferroni-corrected *t* values of 3.04 and 3.35 are required for familywise alpha levels of .05 and .01, respectively.

Table 4
Comparison of French and American Intelligence on Separate WISC-R Subtests for the Total Samples (Ages 6-16), Entering French Scores into American Norms--ADJUSTED FOR GENERATIONAL CHANGES IN THE U. S.

Subtest	American Mean	American SD	French Mean	French SD	Mean Diff.	*t* of Diff.
Verbal						
Similarities	9.95	3.11	8.96	3.07	+0.99	8.57**
Arithmetic	10.11	2.89	9.89	3.21	+0.22	1.97
Comprehension	10.05	2.91	9.49	3.06	+0.56	5.07**
Digit Span	9.95	3.04	8.53	2.98	+1.42	12.60**
Performance						
Picture Completion	10.11	3.00	9.49	3.14	+0.62	5.45**
Picture Arrangement	10.04	3.05	9.44	3.05	+0.60	5.27**
Block Design	10.03	3.03	10.92	3.17	-0.89	-7.75**
Object Assembly	10.02	3.11	10.30	3.21	-0.28	-2.38
Coding	10.02	3.06	9.68	3.01	+0.34	3.00*
Mazes	10.10	3.17	10.42	3.13	-0.32	-2.72

**$p < .01$

Note. Diff. = Difference. The values shown are means for ages 6 to 16 years. Total sample sizes are 2,200 American and 1,066 French children. Bonferroni-corrected *t* values of 2.97 and 3.29 are required for familywise alpha levels of .05 and .01, respectively.

Table 5 reports the Verbal-Performance IQ discrepancy for French children relative to American children, by age, computed from the mean French IQs reported in Tables 1 and 2. The average discrepancy is about 5 points ± 2, with slightly larger differentials observed for ages 6 to 8 (6-8 points) than for ages 9 to 16 (3-5 points).

The reanalysis of IQ differences, by entering mean American scores into French norms and correcting the values for generational changes in France, is summarized in Table 6. The table indicates that Americans averaged about 104 on the Verbal Scale, 101 on the Performance Scale, and 103 on the Full Scale. As in the previous analysis, American children earned higher IQs than French children on the Verbal and Full Scales. French children had a P > V profile of about 5 points relative to the American norms. Consistent with that result is the analogous V > P profile for Americans relative to French norms. However, the P > V discrepancy of about five points for the French reduced to a V > P profile of less than three points for the American. That reduction is a direct result of the differential changes in the French norms during the nearly 6½ years that separated the two standardizations. French generational changes involve substantially greater gains on Performance IQ than Verbal IQ, such that the adjustments made to the American IQs serve to reduce the relative V-P IQ discrepancy between the cultures.

Table 5
French Verbal-Performance IQ Difference, by Age, Relative to American Norms--ADJUSTED FOR GENERATIONAL CHANGES IN THE U. S.

Age	Mean V-P IQ Difference
6.5	-6.15
7.5	-7.86
8.5	-6.55
9.5	-3.95
10.5	-3.41
11.5	-5.51
12.5	-4.87
13.5	-5.23
14.5	-4.69
15.5	-3.48
16.5	-3.14
MEAN	-4.98

Note. These mean V-P IQ discrepancies are computed from the mean French IQs reported in Tables 1 and 2.

Table 6
Comparison of French and American Intelligence on the WISC-R IQ Scales, Entering Mean American Scores into French Norms--ADJUSTED FOR GENERATIONAL CHANGES IN FRANCE

Age	Mean V-IQ	Mean P-IQ	Mean FS-IQ	V-P IQ Difference
6.5	106.09	100.06	103.68	+6.03
7.5	104.84	99.53	102.16	+5.31
8.5	104.21	99.76	101.89	+4.45
9.5	103.16	100.90	101.95	+2.26
10.5	100.50	100.87	100.83	-0.37
11.5	105.04	101.48	103.83	+3.56
12.5	104.41	100.70	102.79	+3.71
13.5	104.62	101.06	103.36	+3.56
14.5	104.41	101.20	103.29	+3.21
15.5	104.52	103.56	104.23	+0.96
16.5	101.32	102.62	101.79	-1.30
MEAN	103.91	101.07	102.71	+2.84

Discussion

The present results indicate that as of the late 1970s, when the second of the two standardizations was conducted, American children scored significantly higher on the WISC-R Verbal Scale, but that no overall differences were observed on the Performance Scale. In fact, however, significant differences emerged within the Performance Scale. French children scored higher on visual-motor measures of spatial ability, most notably on Block Design. From Rapaport's interpretive system (Mayman, Schafer, & Rapaport, 1951), the French children had a strength, relative to American children, on tests of visual organization with essential motor activity (Block Design, Object Assembly, Mazes), but had a relative weakness on tests classified as visual organization without essential motor activity (Picture Completion, Picture Arrangement). Also, Picture Arrangement, Coding, and Digit Span all require auditory or visual sequencing ability, a weakness for French children relative to American children.

From Bannatyne's (1971, 1974) classification of WISC-R subtests, French children's performance on the separate subtests can be meaningfully compared to American norms on Verbal Conceptualization (Similarities, Comprehension), Spatial (Picture Completion, Block Design, Object Assembly), and Sequential (Arithmetic, Digit Span, Coding). (Although Vocabulary is also a Verbal Conceptualization subtest, it had to be excluded from the present analyses.) Using formulas for Wechsler composites provided by Tellegen and Briggs (1967), mean standard scores for French children relative to American norms are as follows: Verbal Conceptualization (95.7), Spatial (101.4), and Sequential (95.8). Relative to American children, French children displayed "Spatial greater than Sequential" and "Spatial greater than Verbal Conceptualization" patterns of about 5½ to 6 points. From a cerebral specialization model (Sperry, 1968; Springer & Deutsch, 1981), French children performed better, relative to American children, on the simultaneous-holistic tasks associated with the processing style of the right hemisphere than on the verbal-sequential tasks believed to be subsumed by the left hemisphere. Again, however, the magnitude of the observed patterns will be less extreme if corrections are made for the differential generational changes of French, instead of American, children.

A cognitive pattern similar to the one observed in this study for school-age children and adolescents on the WISC-R was noted by Quintin-Ercilia (1985) for 100 young French-speaking children from Quebec on the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967). The P > V pattern among those preschool and primary-grade children was so common on the French-translated WPPSI when American norms were applied that it appeared to the researcher to be "normal." And, as indicated previously, patterns similar to the ones seen in this study for French children have been found to characterize Japanese children (Ishikuma et al., 1988; Lynn & Hampson, 1986b, 1986c; Kaufman et al., 1989; Stevenson et al., 1985).

In the present study, the data for the Performance Scale are more reliable and generalizable than the data for the Verbal Scale. All six Performance subtests were translated directly and only a single item was changed (on Picture Completion). Also, with the possible exception of Picture Arrangement, differences observed on the Performance subtests are likely to reflect cognitive differences rather than social or cultural differences. Verbal Scale data must be considered tentative because (a) IQs were prorated from four subtests, one of which (Digit Span) is a supplementary task; (b) Information and Vocabulary, typically the best measures of the general factor and the Verbal Comprehension factor on any Wechsler battery, had to be excluded from the study because they were changed extensively in the French WISC-R; (c) on the 17-item Comprehension subtest, two moderately difficult French items (#s 8 and 10) were substituted for two difficult American items (#s 14 and 16), indicating that the overall item sets were not of comparable difficulty; (d) on the socially-oriented

Comprehension subtest, even identical items may be affected by cultural, as well as cognitive, differences; (e) although the identical Digit Span items were translated into French, test scores can vary as a function of the *linguistic* properties of the number labels (cinq-sept-quatre versus five-seven-four), not just as a function of a child's short-term memory; (f) the higher scores by Americans on the Verbal Scale may reflect a subtle genetic-environmental interaction effect owing to the fact that the items were developed specifically for American culture and not French culture (cf. Harrington, 1975); and (g) the results of the analyses of verbal items are conceivably dependent on the specific content of the items, which means that the present results for the WISC-R may not generalize to the latest American edition of the WISC-R, the Wechsler Intelligence Scale for Children--Third Edition (WISC-III; Wechsler, 1991), because of modifications made to one-third of the items on the Verbal Scale. Regarding the latter point, it is less likely that the use of the WISC-R instead of the WISC-III in this study would affect very much the results for the Performance subtests. For example, the French children's superiority on Block Design is not logically dependent on the specific designs to copy. For similar reasons, the interactions noted by Harrington (1975) are more likely to affect verbal than nonverbal items (with the exception of the culture-loaded Picture Arrangement subtest).

Despite these caveats, one finding suggests that French children may, indeed, have demonstrated less developed verbal intellectual abilities. Their mean of 8.96 on Similarities was .33 *SD* lower than the American mean, a discrepancy that was nearly identical to the .35 *SD* difference on Verbal IQ. Similarities items involve common verbal concepts like "salt" and "lake," that are not likely to be altered by translation; further, the task, though a crystallized skill (Horn, 1985, 1989), is not culture loaded or particularly affected by cultural differences. The verbal conceptualization and reasoning required for success are undoubtedly reflected in the mean cross-cultural difference that was observed. Although one French item was substituted for an American item (METRE-KILO for POUND-YARD), that change is more of a modification than a substitution and is not likely to affect item difficulty. Whereas the differences on Similarities may be subtest-specific and not generalizable to the Verbal Scale as a whole, the differences are, nonetheless, as meaningful to interpret as are the significant discrepancies observed on several Performance subtests.

One of the important inferences from the present study is the degree to which cross-cultural differences are affected by the time interval between the data collection in the two countries. Generational changes have a key impact on data interpretation, and the choice of which culture to correct for can materially affect the result. With the present data set, French children displayed a relative P > V profile when the values were adjusted for generational changes for American children, of 5.0 points, but that difference was only 2.8 points when adjustments were based on generational changes in France.

Flynn's (1987) research suggests that the data on which the American generational changes are based are more valid than the French data. He assigned four categories to the data sets: 1 = verified evidence of IQ gains, 2 = probable evidence, 3 = tentative evidence, 4 = speculative evidence. Flynn assigned a rating of 2 to the American data for children and adolescents, but ratings of 3 and 4, respectively, to the two French data sets that were averaged to yield the generational adjustments used in the present study. The French data that were based on the WISC and WISC-R received the rating of 4. The adjustments for generational changes in the U. S. are, therefore, more justifiable in this study.

Nonetheless, the two separate French data sets studied by Flynn each indicated substantial gains in Full Scale IQ: 6.9 points per decade in the sample tested on the Raven and a verbal-math test, and 3.7 points per decade on the WISC and WISC-R (Flynn, 1987; Kaufman, 1990). Also, each sample gained 5 to 6 points more on nonverbal than verbal tests. In the Raven study the gains were 10.0 points nonverbal and 3.7 points verbal; in the Wechsler study the values were 6.3 and 1.0.

If the average overall gain of 5.3 points per decade is accurate, relative to the gain of 3.0 points for American children, then a second important inference from this study is the tentativeness and datedness of cross-cultural comparisons. If the Full Scale IQ difference of 2.36 points (Table 4) is valid for the late 1970s, then the supposed greater gain for French than American children of 2.3 points per decade means that the advantage would have been wiped out by the late 1980s and the differential might reverse during the 1990s.

At the same time, the P > V profile of French children relative to American children should have increased considerably since the late 1970s. American children's IQs are believed to increase 2.7 points per decade on the Verbal Scale and 3.3 points per decade on the Performance Scale--a net gain in Performance IQ of 0.6 points per decade. In contrast, French children are believed to average a verbal gain of 2.35 points per decade and a nonverbal increment of 8.15 points per decade, a net P > V profile of 5.8 points. Relative to Americans, therefore, French children may be increasing their P > V pattern by 5 points per decade.

All of these data-based statements are obviously conjectural, but they are not far-fetched. Data from several nations in Flynn's study received ratings of 1, such as the 5.8 point per decade gain for Belgium (7.6 points nonverbal, 4.1 points verbal) and the 3.2 point per decade gain for Norway (4.2 points nonverbal, 2.2 points verbal) (Kaufman, 1990, Table 2.4). Cross-cultural comparisons involving Belgium and Norway would likely yield different results each time such a study was undertaken.

These inferences assume that the generational changes are stable over time, which may or may not be the case. Lynn and Hampson (1986a) demonstrated that generational changes for Japanese individuals were substantially larger just after the industrialization following World War II than during the period after 1960. In contrast, the gain of 3 points per decade within the U. S. has been a virtual constant for about 60 years, and is still evident in data obtained on tests normed in the late 1980s and 1990s (Kaufman, 1994; Kaufman & Kaufman, 1993; Wechsler, 1991).

Differential generational changes for nations, as well as possible differences for a nation from one generation to the next, may provide partial explanations for contradictory results in the literature. For example, Ishikuma (1990) failed to detect a High Simultaneous/Low Sequential profile for Japanese children, relative to American children, even though that pattern was cross-validated for several different samples tested on a variety of measures about one to two decades previously (Ishikuma et al., 1988; Lynn & Hampson, 1986b, 1986c; Kaufman et al., 1989; Stevenson et al., 1985).

Just as the results of the present study can only be interpreted in the context of Flynn's research, it is apparent that the differential generational changes across nations--and across scales (verbal or nonverbal) within nations--will affect interpretation of any cross-cultural research that is conducted. Necessarily, all results of such research can only be interpreted as meaningful for the year in which the data were collected. Ideally, differences can be interpreted with reference to factors that are unique to a nation. For example, in France, according to the French Department of Education, about 35% of children at age 2 are in public school, and the percentage rises to 99% at age 3. Day care is not an option. The emphasis in public schools for very young children is often on nonverbal activities such as puzzles and blocks. That variable might potentially be related to their good performance on Block Design and similar visual-motor tasks, and to the much greater generational gain for French children in nonverbal than verbal abilities.

To truly understand cross-cultural differences, however, more direct research is needed. For example, observations and categorizations of French adolescents' strategies for solving Block Design items have been investigated (Beuscart-Zephir & Beuscart, 1988), as have the problem-solving approaches applied by American college students to the Block Design task (Schorr, Bower, & Kiernan, 1982). The Block Design strategies used by comparable samples of French and American children and adolescents should be investigated to better understand the cross-cultural differences observed. Similar studies should be designed to contrast strategies used to solve items on other subtests that produced significant differences, such as Similarities, Digit Span, and Picture Arrangement.

References

Bannatyne, A. (1971). *Language, reading, and learning disabilities.* Springfield, IL: Charles C. Thomas.

Bannatyne, A. (1974). Diagnosis: A note on recategorization of the WISC scaled scores. *Journal of Learning Disabilities, 7*, 272-274.

Beuscart-Zephir, M. & Beuscart, R. (1988). Tests de performances: Une méthode d'analyse des stratégies de résolution. Un example: Le test de cubes du WISC-R. *European Journal of Psychology of Education, 3*, 33-51.

Flynn, J. R. (1983). Now the great augmentation of the American IQ. *Nature, 301*, 655.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains, 1932-78. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Harrington, G. M. (1975). Intelligence tests may favour the majority groups in a population. *Nature, 258*, 708-709.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 267-300). New York: Wiley.

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 61-116). New York: W. H. Freeman.

Ishikuma, T. (1990). *A cross-cultural study of Japanese and American children's intelligence from a* *sequential-simultaneous perspective.* Unpublished doctoral dissertation, University of Alabama.

Ishikuma, T., Moon, S., & Kaufman, A. S. (1988). Sequential-simultaneous analysis of Japanese children's performance on the Japanese McCarthy scales. *Perceptual and Motor Skills, 66*, 355-362.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children interpretive manual.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test (KAIT).* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., McLean, J. E., Ishikuma, T., & Moon, S. (1989). Integration of the literature on the intelligence of Japanese children and analysis of the data from a sequential-simultaneous perspective. *School Psychology International, 10*, 173-183.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature, 297*, 222.

Lynn, R. (1983). IQ in Japan and the United States. *Nature, 306*, 292.

Lynn, R., & Hampson, S. (1986a). The rise of national intelligence: Evidence from Britain, Japan, and the USA. *Personality and Individual Differences, 7*, 23-32.

Lynn, R., & Hampson, S. (1986b). Intellectual abilities of Japanese children: An assessment of 2-1/2-8-1/2-year-olds derived from the McCarthy Scales of Children's Abilities. *Intelligence, 10*, 41-58.

Lynn, R., & Hampson, S. (1986c). The structure of Japanese abilities: An analysis in terms of the hierarchical model of intelligence. *Current Psychological Research and Review, 4*, 309-322.

Mayman, M., Schafer, R., & Rapaport, D. (1951). Interpretation of the WAIS in personality appraisal. In H. H. Anderson & G. L. Anderson (Eds.), *An introduction to projective techniques* (pp. 541-580). New York: Prentice-Hall.

Moon, S. (1988). *A cross-cultural validity study of the Kaufman Assessment Battery for Children.* Unpublished doctoral dissertation, University of Alabama.

Moon, S., Byun, C., McLean, J. E., & Kaufman, A. S. (1994). Sequential simultaneous profile analysis of Korean children's performance on the Kaufman Assessment Battery for Children (K-ABC). *Research in the Schools, 1*, 29-35.

Quintin-Ercilia, P. (1985). Note sur l'usage clinique du WPPSI. *Canadian Psychology, 26*, 214-218.

Schorr, D., Bower, G. H., & Kiernan, R. (1982). Stimulus variables in the Block Design task. *Journal of Consulting and Clinical Psychology, 50*, 479-487.

Sperry, R. W. (1968). Hemispheric deconnection and unity in conscious awareness. *American Psychologist, 23*, 723-733.

Springer, S. P., & Deutsch, G. (1981). *Left brain right brain.* San Francisco, CA: Freeman.

Stevenson, H. W., & Azuma, H. (1983). IQ in Japan and the United States. *Nature, 306*, 291-292.

Stevenson, H. W., Stigler, H. W., Lee, S., Lucker, G. W., Kitamura, S., & Hsu, C. (1985). Cognitive performance and academic achievement of Japanese, Chinese, and American children. *Child Development, 56*, 718-734.

Tellegen, A., & Briggs, P. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology, 31*, 499-506.

Vining, D. R. (1983). Mean IQ differences in Japan and the United States. *Nature, 301*, 738.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence (WPPSI).* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children--Revised (WISC-R).* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1981). *Manuel: Echelle d'Intelligence de Wechsler pour Enfants, Forme Révisée (WISC-R).* Paris: Les Editions du Centre de Psychologie Appliqué (ECPA).

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children--Third Edition (WISC-III).* San Antonio, TX: The Psychological Corporation.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 35-43

The Prediction of Academic Achievement Using Non-Academic Variables

Susan E. Britt
Rutherford County Board of Education

Jwa K. Kim
Middle Tennessee State University

An attempt was made to develop a comprehensive model for predicting academic achievement with non-academic factors by utilizing Structural Equation Modeling. One hundred and forty-seven students enrolled in undergraduate statistics classes served as voluntary participants and were administered a College Achievement Questionnaire and an Academic Self-Concept Scale. A correlation matrix was computed for the total sample and models were tested using the CALIS program. The goodness-of-fit index, adjusted goodness-of-fit, root mean square residual, and chi-square for each model were obtained. Three models were proposed for testing. The Direct Model, the Bio-Model, and the Family Mediated Model. Structural Equation Modeling revealed that both the Bio-Model and the Family Mediated Model serve to predict academic achievement moderately well.

There have been many studies investigating variables that influence academic achievement in an effort to accurately predict future performance. Both academic and nonacademic variables have been assessed across age levels and gender lines. Numerous studies have shown that there is a significant relationship among many of the variables of interest, such as socioeconomic status, gender, study habits, employment, and involvement in social activities. However, there has been a lack of attention to a comprehensive model.

In predicting academic success, researchers have focused mainly on groups of related predictors (Rotter, 1988) or single predictors such as SAT scores or study time (Corley, Goodjoin, & York, 1991; Dickinson & O'Connell, 1990; Dreher & Singer, 1985). This study proposes a comprehensive model that yields a more adequate prediction equation. This model takes five broad factors into account in the prediction process. Included are biological or demographic variables, background and family influence variables, psychological variables, social variables, and academic variables.

Factors that have been found to be predictive of aptitude scores include age bias (Zeidner, 1987), time resources (Bee & Ronaghy, 1990), GPA (Young, 1991), and gender (McCornack & McLeod, 1988). McCornack and McLeod suggested that we need to use separate prediction equations with different slopes and intercepts for men and women. This analysis combines gender and age to form a common factor to be named the "Bio-factor."

Parental expectations and attitudes have been shown to affect children's academic achievements. Crandall, Dewey, Katkovsky, and Preston (1964) found a cross-sex influence of parental behaviors on children's performance. Seginer (1983) suggested that parents' expectations of their children might be both a cause and an effect of academic achievement. Jay and D'Augelli (1991) found a significant difference between African-Americans and Whites with regard to perceived support from friends and family, however, when family income was controlled there was no difference. For this study the family background factors of parents' level of educational attainment and parents' occupation were combined with students' perceptions of parental support and encouragement to form the "Family Factor."

Gadzella and Williamson (1984) found that while a positive self-concept was related to school success, self-concept correlated highly with a measure of study skills. The skills most related to good self-concept and achievement appear to be oral communication and interpersonal relation skills. In another study, Gadzella, Williamson, and Ginther (1985) looked at the relationship between self-concept, locus of control, and academic performance and found that most self-concept subscales had significant positive correlations with the Internal Locus of Control scale. Song and Hattie (1985) found that of 11 facets of self-concept the academic self-concept has the greatest relationship to academic achievement. Mboya (1989) found academic self-concept to be strongly

---

Susan E. Britt is a school psychologist with the Rutherford County Board of Education, Murfreesboro, Tennessee. Jwa K. Kim is an Associate Professor of Psychology at Middle Tennessee State University. Please address correspondence regarding the article to Jwa K. Kim, Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132.

correlated with academic achievement as measured by the California Achievement Test scores.

Data from a study by Uguroglu and Walberg (1986) suggested that motivation by itself would predict achievement. Motivation is apparently multidimensional and interacts with home, social, and peer factors and will be assumed to include variables of time spent studying and tendency to miss classes. Academic self-concept and motivation will combine to form the "Psychological Factor" for this study. These variables are major components revealing one's attitude and ambitions for success.

The "Social Factor" will include those variables which pertain to entertainment and outside employment situations. Green and Jaquess (1987) found that academic performance of high school juniors was not significantly affected by part-time employment. Ganz and Ganz (1988) also found that the number of hours that a student worked each week had no effect on his or her final grade average. Students who did not work at all and those who worked forty or more hours per week received about the same average grade. Television viewing has long been thought of as an instrument that takes time away from studying. Also of interest are the number of visitations that occur during an average week (both the number of times a student goes out to visit and the number of visitors he or she entertains) as well as the number of people that the student considers to be good friends.

Variables which are generally included in research dealing with student achievement include SAT and ACT scores, study habits, and class attendance. For this study, past performance and performance on standardized achievement tests will constitute the "Academic Factor."

This study tested proposed models in an effort to explore inter-relationships that combine to affect student levels of academic performance. Based on the five latent factors, three models were proposed. The first was the "Direct Model" which would be expected to demonstrate a relatively equal influence of each of the factors on academic achievement. In this model biological and psychological variables, family background, social experiences, and academic factors would each have a significant impact on academic achievement. Second was the "Bio-based Model" suggesting that biological variables are the fundamental cause of all other factors including academic achievement. The third, the "Family Mediated Model," postulates a dynamic interaction of factors where family background variables mediate the impact of other factors all of which are influenced by biological variables.

This study investigated the following hypotheses:

1. The Direct Model will be a significant model for the prediction of academic achievement.
2. The Bio-based Model will be a significant model for the prediction of academic achievement.
3. The Family Mediated Model will be a significant model for the prediction of academic achievement.

Method

Participants

One hundred and forty-seven students who were enrolled in undergraduate statistics classes at a southeastern university served as participants for this study. Of these participants, 89 were female, and 58 were male. Ethnic representation consisted of 136 Whites, 9 African-Americans, and 2 Asians. The participants ranged in age from 19 to 56 years. The participants volunteered to participate and were at minimal risk of harm, in accordance with the "Ethical Principles of Psychologists" (American Psychological Association, 1992). The participants' mean age and standard deviation were 24.5 years and 6.34 years respectively. The mean and standard deviation high school GPAs were 3.1 and .49 respectively.

Materials

Each participant was administered a two-part questionnaire (see Appendix). The first part of the questionnaire contained items to collect demographic data, academic history, and patterns of social activities. The second part of the questionnaire consisted of the Academic Self-Concept Scale (Reynolds, 1988). This scale was developed as a measure of an academic facet of general self-concept in college students. The validity data suggest that the Academic Self-Concept Scale measures a facet of self-concept specific to an academic self-attitude and is an academic rather than an aptitude dimension of self-concept (Reynolds, Ramirez, Magrina, & Allen, 1980). On the basis of responses from 427 college students, the final form of the Academic Self-Concept Scale has an internal consistency reliability of .91. Validity was established by correlating the ASCS with grade point averages of students and with their scores on the Rosenberg Self-Esteem Scale. A multiple regression analysis of the ASCS with GPA and Rosenberg scores as predictor variables resulted in a multiple correlation of .64.

Procedure

Each participant was administered the two-part pencil and paper questionnaire during the regularly scheduled statistics class. The questionnaire took approximately 30 minutes to complete. Each student signed a consent form, and complete anonymity was maintained. A correlation matrix was computed for the total sample, and models were tested using the CALIS program in SAS (SAS Institute, Inc., 1990). The goodness-of-fit index (GFI) and adjusted GFI, root mean square residual (RMR), and chi-square for each model were obtained.

Factors were determined through consideration of the findings in the research literature concerning academic achievement. For these analyses the original models were retained even though the contributions of some variables to corresponding factors were less than satisfactory.

Results

The intercorrelations, means, and standard deviations for the variables used in the present study are presented in Tables 1 and 2. The intercorrelations among the variables of interest were generally low to moderate ($r = .20\text{-}.50$).

Table 1
Intercorrelations for Variables Included in the Prediction Models

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. GEN	-	-14	-05	16	-16	-03	-37	-00	-20	05	07	16	27	23	28	-08	08	13	15	00	04	-02	14
2. AGE		-	-07	-31	34	12	13	-06	20	-07	-10	-23	-32	-15	-11	28	-09	-16	-05	01	-25	27	28
3. ETH			-	03	05	-01	06	-12	-19	-01	13	02	-05	-07	-08	07	-16	-13	-10	-19	-06	04	04
4. SKP				-	-26	-33	-23	-04	-24	-06	17	33	26	16	-01	-10	00	15	01	-02	-02	02	07
5. STY					-	22	19	-23	08	-12	-01	-08	-16	-08	01	18	-11	-23	03	04	03	-12	-16
6. ASC						-	27	07	29	-03	-07	-14	04	06	03	05	04	-19	09	03	01	09	11
7. GPA							-	27	24	10	-16	-26	-19	-01	-14	11	-12	-31	-12	-10	06	03	-12
8. ACT								-	19	21	-02	-17	03	-07	-07	-10	15	-01	14	07	02	-01	06
9. STS									-	-04	-12	-17	05	02	-04	00	02	-04	05	08	-23	-14	-11
10. WRK										-	-04	-18	-12	-16	01	-04	-01	02	-09	-21	-11	-08	-05
11. TV											-	08	-01	-06	02	06	-02	04	01	10	24	15	22
12. PTY												-	35	38	14	-16	19	19	07	08	16	10	10
13. GOV													-	55	30	-13	15	15	10	02	08	09	21
14. VIS														-	38	-13	02	12	09	13	11	11	12
15. FRD															-	-10	02	13	12	15	-04	-09	-00
16. SIB																-	-17	-27	-22	-12	-19	-30	-32
17. MED																	-	46	53	34	28	22	25
18. MJB																		-	28	31	27	25	28
19. DED																			-	58	12	21	19
20. DJB																				-	15	08	04
21. INV																					-	58	55
22. ENC																						-	74
23. SUP																							-

Note: All decimal points have been dropped. Gender (GEN); Ethnicity (ETH); Skips Class (SKP); Hours Study (STY); Academic Self-Concept (ASC); High School GPA (GPA); ACT Score (ACT); Statistics Points (STS); Hours Work (WRK); Hours Watch TV (TV); Number Parties per Month (PTY); Social Visits per Week (GOV); Number of Visitors per Week (VIS); Number of Good Friends (FRD); Number of Siblings (SIB); Mother's Educational Level (MED); Mother's Occupation (MJB); Father's Educational Level (DED); Father's Occupation (DJB); Parental Involvement in School Activities (INV); Encouragement to Attend College (ENC); Parental Support of Educational Goals (SUP). $p < .05$ if coefficient $> .17$; $p < .01$ if coefficient $> .22$.

The first model tested was the Direct Model which hypothesized that each of the factors would have a relatively equal influence on academic achievement. This hypothesis was not supported. The analysis revealed a goodness-of-fit index (GFI) of .8230; a GFI adjusted for degrees of freedom (AGFI) of .7981; a root mean square residual (RMR) of .1473; a $\chi^2(242, N = 147) = 282.97$, $p = .03$. The Direct Model does not fit the correlation matrix, indicating an inadequate model for the prediction of academic achievement.

BEST COPY AVAILABLE

Table 2
Means and Standard Deviations for Variables

Factor	Variable	Mean	Standard Deviation
BIOLOGICAL			
	Age	24.5	6.34
PSYCHOLOGICAL			
	Skip Classes	1.8	.81
	Hours Study	2.2	.92
	ASC	114.6	12.62
ACADEMIC			
	High School GPA	3.1	49
	ACT Score	21.4	3.79
SOCIAL			
	Hours Work/Week	21.3	13.77
	Hours Watch TV	2.5	1.65
	Parties/Month	1.7	2.29
	Visits/Week	2.3	2.00
	Visitors/Week	3.4	5.73
	Number Friends	8.0	10.85
FAMILY			
	Number Siblings	2.1	1.50
	Involvement	6.6	2.84
	Encouragement	8.4	.57
	Support	8.7	2.24

The Bio-Model is presented in Figure 1 and shows standardized path coefficients and R^2-values obtained for each factor. A path coefficient indicates the size of the direct effect the exogenous variable has on the endogenous variable. The amount of variance (the R^2-values) accounted for in each endogenous variable by the exogenous variable or variables having a direct effect on it appears in parentheses.

Structural Equation Modeling produced a GFI of .8135; an AGFI of .8065; a RMR of .1508; a $\chi^2(266, N = 147) = 298.132$, $p = .08$. The Bio-Model fits the correlation matrix moderately well.

This model showed that 99% of the variance in the Academic factor was accounted for by a combination of the Biological, Social, Psychological, and Family factors. Almost 72% of the variance in the Social factor, 68% of the variance in the Psychological factor, and 13% of the variance in the Family factor was accounted for by the Biological factor. The results also showed that the Biological factor was a significant predictor for the Social, Psychological, and Family factors. Furthermore, the combination of the Biological and Psychological factors was a significant predictor of the Academic factor.

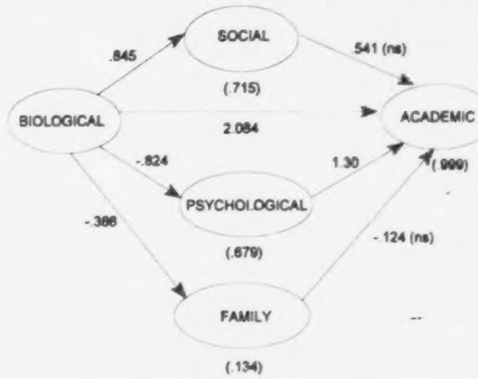


Figure 1. Bio-Based Model. Model of the relationships between the latent variables affecting academic achievement when Biological variables directly impact all other factors.

The final model tested was the Family Mediated Model which is presented in Figure 2. This model resulted in a GFI of .8294; an AGFI of .8112; a RMR of .1527; a $\chi^2(264, N = 147) = 288.633$, $p = .14$. The Family Mediation Model resulted in the highest goodness-of-fit index combined with the greatest probability for the chi-square which indicates that this model fits the correlation matrix moderately well also and only slightly better than the Bio-Model.

This model showed that 100% of the variance in the Academic factor was accounted for by the Biological, Psychological, Social, and Family factors. As well, 99% of the variance in the Psychological factor and 61% of the variance in the Social factor is accounted for by a combination of the Biological and Family factors while just over 11% of the variance in the Family factor was accounted for by the Biological factor.

The Family Mediated Model showed that the Biological factor was again a significant predictor for the Social, Psychological, and Family factors and that the combination of the Biological factor and the Psychological factor was a significant predictor for the Academic factor.

A chi-square test for comparing the two models was performed and results indicate that the Bio-Model and the Family Mediated Models were significantly different, $\chi^2(2, N = 2) = 9.50$, $p < .05$. According to this chi-square test the Family Mediated Model fits the correlation matrix slightly better than the Bio-Model.

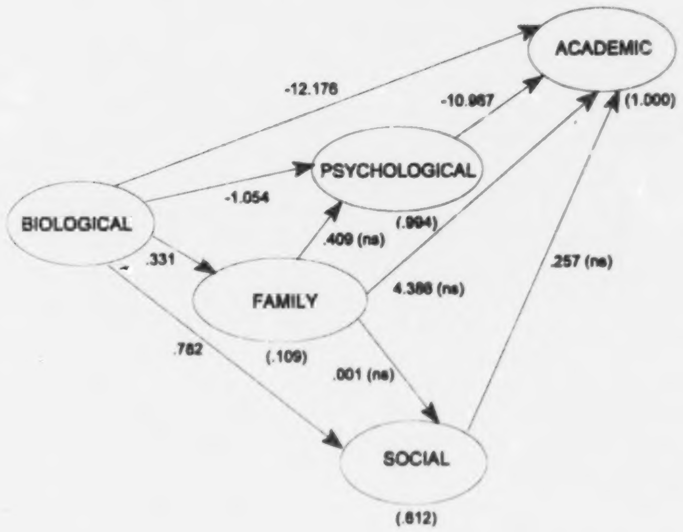


Figure 2. Family Mediated Model. Model of the relationships between latent variables when Family variables mediate the influence of other factors.

Discussion

The Bio-Model fits the correlational matrix moderately well with a relatively high GFI. According to this model, Biological, Social, and Psychological factors are all significant predictors for the Academic factor. The Biological factor is a significant predictor for Social, Psychological, and Family factors.

One interesting phenomenon is that the Biological factor inversely predicts the Psychological and Family factors. Gender is one factor which could influence this relationship. In the original data, females (recorded as 1) report higher levels of academic motivation than males (recorded as 2). Although the magnitude is not large, gender may also account for the inverse relationship to the Family factor. The data showed that females perceived higher levels of family support than males.

The Family Mediated Model showed a moderate goodness-of-fit index for the Academic factor. The Biological factor is the exogenous variable which contributes significantly to the Family, Social, Psychological and Academic factors. Both the Family and Social factors are non-significant predictors for the Academic factor. The Psychological factor is inversely predicting the Academic Achievement factor. This inverse relationship is expected to be a result of the same factors as occurred in the Bio-Model.

Although the statistical tests showed that the Family Mediated Model and the Bio-Model fit the correlation matrix moderately well, there are additional aspects to consider. One is the unlikely R^2 which resulted to explain the Academic factor of the two significant models. It is unusual that we could account for 99% of the variance in the Academic factor of the Bio-Model or for 100% of the variance in the Academic factor of the Family Mediated

Model. A better estimate of the endogenous variables is necessary. One cause of the abnormal variance has to do with the estimated variance of the Academic factor being very large, 10,692 from the Family Mediated Model and 23,394 from the Bio-Model. In addition the absolute values of some coefficients were very large (10.967 and 12.176).

A second problem has to do with the instability of the prediction coefficient for the Academic factor using the combination of the Biological, Social, Psychological, and Family factors. Although some adjustment was made, there were still unknown factors contributing to this instability.

Possible explanations of the above abnormalities include an insufficient sample size ($N = 147$). The present study's target population consisted of students enrolled in undergraduate statistics classes for the social sciences at a regional university. An anticipated sample size of 200 to 250 was reduced through simple attrition which normally occurs in these classes.

Another possibility is the inaccurate direction of the path coefficient. For example, in the Family Mediated Model there are numerous possibilities for other, more accurate paths. It may be that Social factors impact on Family which in turn might have a greater impact on Psychological factors. Or, Psychological factors may be a more important contributor to Family or Social factors than Biological factors. Many different paths may produce very different and more accurate models to use in the prediction of achievement.

A possible weakness in this study has to do with the Family Income item on the College Achievement Questionnaire. This item was deleted from the analysis because it was asked in such a way as to be interpreted differently by each student. Although it is believed that income has a significant impact on many factors in a college student's life, the information must be gathered so as to be comparable across individuals. Also, there were too few minorities represented to give the ethnicity item the variability needed to make a significant contribution.

Again, it should be taken into account that this is the first attempt to develop a comprehensive model for the prediction of academic achievement. These models were based on theoretical speculation and the scientific literature. There are other variables that need to be considered for inclusion such as career goals and income goals. The study could broaden its target population by including students in general introductory psychology classes as well as students from multiple university settings. Individual variables that make up each factor should be analyzed for their contributive properties. Preliminary analysis through multiple regression and exploratory factor analysis might be performed to produce more relevant factors. In conclusion, although both the Bio-Model and the Family Mediated Model fit the correlation matrix moderately well, the Family Mediated Model is the more stable of the two.

References

American Psychological Association. (1992). Ethical principles of psychologists. *American Psychologist, 47*, 1597-1611.

Bee, R. H., & Ronaghy, H. A. (1990). A time budget analysis of collegiate majors. *College Student Journal, 24*, 72-77.

Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw Hill.

Corley, E. R., Goodjoin, R., & York, S. (1991, February/March). Differences in grades and SAT scores among minority college students from urban and rural environments. *The High School Journal*, 173-177.

Crandall, V., Dewey, R., Katkovsky, W., & Preston, A. (1964). Parents' attitudes and behaviors and grade-school children's academic achievements. *Journal of Genetic Psychology, 104*, 53-66.

Dickinson, D. J., & O'Connell, D. Q. (1990). Effect of quality and quantity of study on student grades. *Journal of Educational Research, 83*, 227-231.

Dreher, M. J., & Singer, H. (1985). Predicting college success: Learning from text, background knowledge, attitude toward school, and the SAT as predictors. In J. A. Niles (Ed.), *Issues in literacy: A research perspective* (pp.362-366). Rochester, NY: National Reading Conference.

Gadzella, B. M., & Williamson, J. D. (1984). Study skills, self-concept, and academic achievement. *Psychological Reports, 54*, 923-929.

Gadzella, B. M., Williamson, J. D., & Ginther, D. W. (1985). Correlations of self-concept with locus of control and academic performance. *Perceptual and Motor Skills, 61*, 639-645.

Ganz, M. N., & Ganz, B. C. (1988). An assessment of some factors which affect grades in the community college. *College Student Journal, 22*, 171-175.

Green, G., & Jaquess, S. N. (1987). The effect of part-time employment on academic achievement. *Journal of Educational Research, 80*, 325-329.

Jay, G. M., & D'Augelli, A. R. (1991). Social support and adjustment to university life: A comparison of African-American and White freshmen. *Journal of Community Psychology, 19*, 95-108.

Mboya, M. M. (1989). The relative importance of global self- concept and self-concept of academic ability in predicting academic achievement. *Adolescence, 24,* 39-45.

McCornack, R. L., & McLeod, M. M. (1988). Gender bias in the prediction of college course performance. *Journal of Educational Measurement, 25,* 321-331.

Reynolds, W. M. (1988). Measurement of academic self-concept in college students. *Journal of Personality Assessment, 52,* 223-240.

Reynolds, W. M., Ramirez, M. P., Magrina, A., & Allen, J. E. (1980). Initial development and validation of the academic self-concept scale. *Educational and Psychological Measurement, 40,* 1013-1016.

Rotter, N. G. (1988). Student attrition in a technological university: Academic lifestyle. *College Student Journal, 22,* 241-248.

SAS Institute, Inc. (1990). *SAS technical report P-200 SAS/STAT software: CALIS and LOGISTIC procedures* [Computer program manual]. Cary, NC: SAS Institute.

Seginer, R. (1983). Parents' educational expectations and children's academic achievements: A literature review. *Merrill-Palmer Quarterly, 29,* 1-23.

Song, I., & Hattie, J. (1985). Relationships between self-concept and achievement. *Journal of Research in Personality, 19,* 365-372.

Uguroglu, M. E., & Walberg, H. J. (1986). Predicting achievement and motivation. *Journal of Research and Development in Education, 19,* 1-12.

Young, J. W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. Journal of Educational Measurement, *28,* 37-47.

Zeidner, M. (1987). Age bias in the predictive validity of scholastic aptitude tests: Some Israeli data. Educational and Psychological Measurement, *47,* 1037-1047.

Appendix
COLLEGE ACHIEVEMENT QUESTIONNAIRE

Social Security # ____________ Gender ______ Age ______

Ethnicity
White ________ African-American ________
Asian ________ Native American ________
Hispanic ________

College Major ____________ High School GPA ________

Year in School
Freshman ________ Senior ________
Sophomore ________ Graduate School ________
Junior ________

ACT Score ________ SAT (Total) Score ________

Hours Per Week Employed (Average) ________

Number of Siblings Brothers ________ Sisters ________

Parents in the Home (Check all that apply)
Mother ________ Step-Parent ________
Father ________ Other ________

Mother's Educational Level
Less than high school graduate ________
High school graduate ________
Some college ________
College graduate ________
Graduate school ________

Mother's Occupation ____________________

Father's Educational Level
Less than high school graduate ________
High school graduate ________
Some college ________
College graduate ________
Graduate school ________

Father's Occupation ____________________

Yearly Family Income (Check One)

Less than $ 5,000 ________ $ 5,000--$ 9,999 ________
$10,000--$14,999 ________ $15,000--$19,999 ________
$20,000--$24,999 ________ $25,000--$29,999 ________
$30,000--$34,999 ________ $35,000--$39,999 ________
$40,000--$44,000 ________ $45,000--$49,999 ________
Over $50,000 ________

On a scale of 0 to 10 (Check One):

How much were your parents involved with your school activities during your elementary through high school years?

0 ____ 1 ____ 2 ____ 3 ____ 4 ____ 5 ____ 6 ____ 7 ____ 8 ____ 9 ____ 10

How much did your parents encourage you to attend college?

0 ____ 1 ____ 2 ____ 3 ____ 4 ____ 5 ____ 6 ____ 7 ____ 8 ____ 9 ____ 10

How supportive do you feel your parents are of your educational goals?

0 ____ 1 ____ 2 ____ 3 ____ 4 ____ 5 ____ 6 ____ 7 ____ 8 ____ 9 ____ 10

How often do you skip classes?

Almost never ________
Seldom (once a month) ________
Occasionally (once a week) ________
Almost always ________

How much time do you spend studying for THIS class (Statistics) per week?

0-1 Hour ________ 3-4 Hours ________
2-3 Hours ________ More than 4 ________

During an average weekday, how many hours do you spend watching television (including movies on the VCR)?

0 ____ 1 ____ 2 ____ 3 ____ 4 ____ 5 ____ 6 ____ 7 ____ 8 ____ 9 ____ 10

How many movies do you watch at the theater per week: ________

How many parties do you attend per month: ________

Number of times you go out visiting per week: ________

Number of visitors to your home (social) per week: ________

How many people do you consider good friends: ________

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 45-50

Testing at Higher Taxonomic Levels: Are We Jeopardizing Reliability By Increasing the Emphasis on Complexity?

Andrea D. Clements
East Tennessee State University

Lori Rothenberg
West Georgia College

Undergraduate educational psychology exams were analyzed for proportions of test items at each level of Bloom's Taxonomy, item format, and test length. Forty-eight usable responses were received from a mailing to 200 randomly selected colleges and universities. Analyses indicated significant relationships between item complexity and test length even when taking into account item format (e.g., $R^2 = .61$, $p < .0001$, partial correlation = $-.36$, $p < .05$). Therefore emphasis on use of higher items may be related to use of shorter tesis, thereby jeopardizing relilability.

What is our greater concern, preparing statistically reliable tests, or measuring whether our students can use the information to which they are exposed in our classes, and how do our testing methods relate to these factors? While both are desirable, there is a consensus among educational theorists that having students participate in higher order thinking should be the main goal of education (Raudenbush, Rowan, & Cheong, 1993), and that there are and will continue to be needs for classifying, analyzing, synthesizing, and applying knowledge (Lapointe, 1984). Lapointe (1984) noted that since the mid 1970's, National Assessment of Educational Progress (NAEP) data showed basic achievement improvements in reading, math, science, and writing, but that higher order thinking skills such as the abilities to infer meaning from a passage, classify and solve mathematics problems, interpret meanings of scientific data, and marshall arguments in support of a thesis, have declined. Lewis and Smith (1993) define higher order thinking as follows:

> Higher order thinking occurs when a person takes new information and information stored in memory and interrelates and/or rearranges and extends this information to achieve a purpose or final possible answer in perplexing situations. (p. 136)

The activities mentioned by Lapointe (1984), and Lewis and Smith (1993) are the types of activities measured by test items or activities classified in the upper levels of Bloom's Cognitive Taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956), namely application, analysis, synthesis, and evaluation.

Before deciding that this trend away from higher level thinking is detrimental, it must be asked whether there is any benefit to requiring students to think at higher levels. If not, why expend the effort involved in generating such measures? Cox (1965) found that items classified at the knowledge, or lowest, level of Bloom's Taxonomy were poor discriminators between students who understood course material and those who did not, but that items at the application level, a higher level, were the best discriminators. Ennis (1993) suggests categorization according to Bloom's Taxonomy, and inclusion of items from its upper levels to be a good way to assess these more complex types of student thinking. Annis and Annis (1987) found that rereading material enhanced performance on knowledge, comprehension, application, and analysis questions, which represent the lower four levels of Bloom's Taxonomy, but did not significantly enhance performance on synthesis and evaluation items, the upper two levels. It would follow that activities

---

Andrea D. Clements is Assistant Professor of Human Development and Learning, College of Education, at East Tennessee State University. Lori Rothenberg is Assistant Professor of Educational Psychology, School of Education, at West Georgia College, but will begin a position as Assistant Professor of Educational Evaluation and Research in the College of Education at Wayne State University in Fall, 1996. Correspondence regarding this article shoud be sent to the first author at the College of Education, East Tennessee State University, Johnson City, TN, 37614, or via email at clements@etsuvax.east-tenn-st.edu.

requiring more complex types of reasoning are necessary to enhance performance at these upper two levels.

Tradition as well as logic suggest that the best way to determine whether our students are able to perform activities that qualify as higher order thinking is to test these activities. Testing with higher level items would require the types of thinking included in the Raudenbush et al. (1993) definition of higher order thinking. Lewis and Smith (1993) point out that it is necessary to present situations or questions that cannot be answered through simple recall in order to truly assess higher order thinking.

Although there is much talk of the need to teach and assess higher order thinking skills and the need to ensure that students are able to solve problems, many of our current assessments require far less sophisticated processes. Teachers' oral questioning techniques or patterns place heavy emphasis on rote recall and memorization (Aiken, 1982; Clevenstine, 1987; Gall, 1970; Rinchuse & Zulio, 1986; Smith, 1984), which would be classified at the lower levels of Bloom's taxonomy, namely knowledge and comprehension. Only about 20% of teachers' questions require students to think, and the others involve recall of material or procedural information (Gall, 1970). In soliciting undergraduate educational psychology instructors "best tests," Clements, Hamilton, and Rothenberg (1994) found that 68% of the 1,515 items received were at the knowledge level, and more than 17% were at the comprehension level. Aiken (1982) noted "(t)he great majority of multiple-choice items in instructors' manuals and other sources of test materials assess only 'simple knowledge or recognitive memory'" (p. 803). Smith (1984) reported that when commercially prepared tests were analyzed, 72% of the 2,689 items assessed were at the knowledge level, 15% at the comprehension level, and less than 2% fell at or above the analysis level. Other sources have noted similar trends when analyzing various test formats (Clevenstine, 1987; Rinchuse & Zulio, 1986).

Understanding that most tests tend to be heavily laden with lower level items, we concurred that a closer look at the characteristics of tests with varying proportions of higher level items was in order. Educational psychology course tests were chosen because test construction is taught as an integral part of educational psychology, and those teaching such courses should have been educated in test construction issues. We decided that this would give us the greatest chance of finding higher proportions of complex items and well constructed tests. The purpose of this investigation was to determine whether there was a significant relationship between test length, and complexity of items, and if found, what implications it would have for testing practice. Another factor that warranted consideration in the investigation of complexity within tests was item format. It is generally accepted that to require a student to synthesize (i.e., produce a new product) one must use a constructed response item. We chose to investigate the strength of the relationship between item complexity and item format and its implications as well. Gronlund (1993) indicated that selected response items (multiple choice, true-false, matching, some fill-in-the-blank) are objective, simple, and highly reliable, while constructed response items, such as essay items, are subjective, difficult, and less reliable. He went on to say that subjective grading and low numbers of items are two of the four main factors which serve to reduce reliability of tests. It was hypothesized that tests which require higher order tasks would be both shorter and contain a higher proportion of subjective items. Popham (1995) supports this position by saying "(t)he most serious problem with essay items...is the difficulty that teachers have in scoring students' responses reliably" (p. 123).

For the purpose of this study, items will be discussed in terms of two dichotomies: (a) item complexity including higher level (upper four levels of Bloom's taxonomy) and lower level (lower two levels of Bloom's taxonomy), and (b) item format including subjective (essay and short answer) and objective (multiple choice, true-false, matching, and fill-in-the-blank). It was hypothesized that there would be a significant relationship between the item complexity and test length when controlling for item format.

Method

Subjects

Two hundred colleges and universities listed in the Peterson's Guide to Four Year Colleges and Universities were randomly selected using a table of random numbers from those that have programs in at least one of the following: Education, Elementary Education, Middle Grades Education, and/or Secondary Education. Of those contacted, responses of some type were received from 56 institutions (28%). Of those, 48 (24%) yielded usable information.

This is a low return rate. However, the fact that the proportions of sizes and selectivity of responding institutions and that the tests provided were similar in characteristics (complexity and item format) to those found in earlier studies lends some support to the representativeness of this sample. Any generalizations should be made cautiously. There is no strong evidence that these tests are representative of college level tests, or even of educational psychology tests. Even after follow-up contact, nothing is known about tests of those institutions which did not respond.

Instrumentation

A letter was sent to each of the 200 institutions requesting information about several institutional characteristics. Of interest were selectivity of the institution, size of the institution, GPA required for admission into a teacher education program, whether teacher education requirements included a course in Educational Psychology, and the average size of their undergraduate Educational Psychology course(s). Institutions were also asked to provide a copy of what the Educational Psychology instructor(s) felt was their "best" undergraduate Educational Psychology test. Each test was analyzed for proportions of levels of Bloom's taxonomy tested and item formats used.

Procedure

Letters requesting demographic data and tests were mailed in January, 1993. Several institutions failed to include tests with their answers to the demographic questions, so individual letters requesting tests were mailed as a follow-up for these. The follow-up letter was mailed in March, 1993. This resulted in an eventual response of some type from 56 institutions. Forty-eight of those responses yielded usable information. The others were not at the listed address. Demographic data were compiled. Thirty-five tests were provided and were used for all additional analyses.

Tests were initially analyzed by two researchers knowledgeable in Bloom's cognitive taxonomy, using identical lists of criteria describing each level of the taxonomy to classify each item's level of complexity. Aggregating all tests resulted in 1,473 items. For those items on which the two initial raters did not agree, a third rater categorized only those items using the same criterion list without consulting the two initial raters. On the 151 items for which all three raters gave different levels, two raters, both assistant professors of educational psychology, discussed the items and came to a consensus for each item using the same criteria used in the initial rating. Multiple regression was used to determine whether there was a significant relationship between test length and item complexity when controlling for item format.

Results

The percentage of total items at each level of Bloom's taxonomy and ranges of proportions of items at each level are given in Table 1. As expected, the largest proportion of items was at the knowledge level, and the lowest proportion was at the evaluation level. The percentage of total items of each format and proportions of items of each format are given in Table 2. The highest proportion, by far, was multiple choice, and the lowest was fill-in-the-blank.

Table 1
Percentage of Items at Each Level of Bloom's Taxonomy for Aggregated Items and By Test

Bloom Level	Overall Percentage (aggregated items)	Lowest Percentage on Any Test	Highest Percentage on Any Test	Mean Percentage, Standard Deviation
Knowledge	58	0	84.87	42.69, 29.16
Comprehension	17.2	0	58.33	20.98, 14.64
Application	17.1	2.56	66.67	20.65, 14.23
Analysis	5.2	0	60	9.97, 14.83
Synthesis	1.5	0	30	3.02, 7.46
Evaluation	0.8	0	37.5	2.68, 7.77

Table 2
Percentage of Items of Each Format for Aggregated Items and By Test

Item Format	Overall Percentage (aggregated items)	Lowest Percentage on Any Test	Highest Percentage on Any Test	Mean Percentage, Standard Deviation
Multiple Choice	75.43	0	100	57.63, 39.67
True - False	6.88	0	50	4.16, 1.08
Matching	3.68	0	27.12	3.22, 8.23
Fill-in-the-blank	2.65	0	52.94	2.21, 9.59
Short Answer	5.11	0	92.9	8.50, 18.75
Essay	6.26	0	100	23.80, 39.32

Test lengths ranged from 1 item to 119 items. The mean number of items was 43 with a standard deviation of 30. The highest percentage of higher level items was 89% on a 9 item test, and the lowest was 7% on a 100 item test. The highest percentage of objective items in any test was 100% ($n = 13$), and there were eight tests which contained only subjective items.

Analyses revealed that there was a significant multiple correlation among the total number of items on tests, the percentage of higher level items on tests, and the percentage of subjective items on tests ($R^2 = .61$, $p < .0001$). In other words, 61% of the variability in the number of items on tests is accounted for by the percentages of higher level items and of subjective items. The partial correlation of test length with percent of higher level items partialling the percent of subjective items was -.36 ($p < .05$). This indicates that the proportion of higher level items is stongly negatively related to test length even when taking the proportion of subjective items into account.

There was a significant multiple correlation among the total number of items on tests, the percentage of lower level items on tests, and the percentage of objective items on tests ($R^2 = .60$, $p < .0001$). This indicates that 60% of the variability in the number of items on tests is accounted for by the percentages of lower level items and of objective items. The partial correlation of test length with percent of lower level items partialling the percent of objective items is .39 ($p < .05$). This indicates that the proportion of lower level items is stongly positively related to test length even when taking the proportion of objective items into account.

Discussion

As hypothesized, there was a significant correlation between test length and complexity when controlling for item format. The relationship was significant regardless of whether lower level and objective items were entered, or higher level and subjective items were entered into the analyses. Before partialling item type, the multiple correlations were quite high, indicating that all three variables were interrelated. However, the relationship remained significant, though weaker than the multiple correlation using three variables, when partialling item type. This is important because there seems to be an actual relationship between item level and test length. Perhaps this is related to the facts that often test banks contain a large proportion of lower level items (Aiken, 1982) increasing access to lower level items, and lower level items seem to be easier to construct. Therefore, to come up with a large number of items (needed for a longer test) one might be more likely to resort to test banks and quickly turned out lower level items.

The fact cannot be ignored that some of the variability in test length is accounted for by item format. One would expect tests primarily containing objective items to be longer because objective items can be completed more quickly by students. An instructor can include more items per test and reasonably expect students to be able to complete them. The inverse would be true of subjective tests. As items require more time to complete, fewer can be included in a test.

The main concerns brought about by our findings relate to reliability. The tests which did have a high proportion of higher level activities required tended to be shorter than those that did not. There is the commonly accepted position that shorter tests carry with them lowered reliability (Cronbach, 1970; Gronlund, 1993). Also, those tests which contain higher proportions of higher level items tended to contain higher proportions of subjectively graded items, another threat to reliability (Gronlund, 1993; Popham, 1995).

If indeed shorter tests are required to allow students time to perform higher level tasks, which appears to be the case from this study, we seem to be forced to choose between requiring higher level tasks and maintaining statistical reliability. This, of course, assumes that classroom tests are reliable in the first place. Needless to say, the chance of having acceptable reliability remains higher with the longer instruments. One suggestion offered by Gronlund (1993) is to test more frequently when time does not permit using longer tests. This, in essence, provides a "longer" instrument administered over multiple sessions. If, indeed, our goal is to have students be able to functionally use material contained in courses, then giving up class time for this functional use would be defensible. This might lessen the incidence of very short tests (e.g., those in this study with fewer than ten items).

As shown in the current study, the strong relationship between test length and complexity is commingled with the strong relationship between item format and test length and item format and complexity. Most of us who strive to test at those upper levels of Bloom's Taxonomy know that this is usually accomplished by using constructed response items such as essay and short answer, which carry their own threats to reliability. Suggestions have been offered to make constructed response item scoring more reliable through the use of item rating check sheets, criteria lists, and use of multiple raters (Gronlund, 1993).

Classroom test instruments are not the only types of measures involved in this complexity versus reliability debate, however. The trade-off of lowered reliability for measurement of higher order thinking skills is a primary concern for the current performance assessment movement as well. In high-stakes testing situations, the main reliability concern is inter-rater reliability. The solution has been to thoroughly train raters. However, in classroom testing situations--where there is only one rater (the professor or teacher)--the issue of reliability is more difficult to resolve. In addition to having to justify possible rater differences, task comparability becomes an issue.

Alternative conceptions of reliability have been suggested in the literature. Moss (1994) argues that the emphasis on "high agreement" reliability detracts from the possibility that a highly valid test may not necessarily be reliable in the traditional sense. She suggests the use of critical dialogue and confrontation as is commonly used in the functioning of a university search committee.

Delandshere and Petrosky (1994) presented a second notion of reliability that involves getting a second expert to confirm the original judgment. This is more likely the way many professors operate. When first teaching a course, professors may consult with colleagues on the appropriateness of assessment devices. Also, when students appeal grades, the reasonableness of the professor's assessment devices may be examined. Although these alternative conceptions of reliability have been proposed, they have their limitations and are not widely accepted.

A last area which necessitates discussion is that any type of higher order testing will take more work. Either more time will be required to write complex items before testing, or more time will be required to grade constructed response items after testing. It remains easier to dash off a knowledge level objective test, or better yet, take one straight from the instructor's manual, but in the interest of benefiting our students we encourage that bit of extra work.

This particular study is beset with several limitations (e.g., low response rate, questionable sample representativeness) which should be addressed before any sweeping generalizations can be made. Tests from other fields within higher education should be analyzed to confirm these findings. Characteristics of tests from K - 12 classes would also be of interest. It seems appropriate to consider other requirements within courses, including written assignments and field-based activities when discussing complexity. It is highly likely that primarily lower level tests are supplemented with at least some of these types of activities.

References

Aiken, L. R. (1982). Writing multiple-choice items to measure higher-order educational objectives. *Educational and Psychological Measurement, 42*, 803-806.

Annis, L. F., & Annis, D. B. (1987, April). *Does practice make perfect? The effects of repetition on student learning.* Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Services No. ED 281 861)

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. Handbook I: Cognitive domain.* New York: David McKay.

Borg, W. R., Gall, J. P., & Gall, M. D. (1993). *Applying educational research: A practical guide.* New York: Longman.

Clements, A. D., Hamilton, N., & Rothenberg, L. (1994, November). *An analysis of undergraduate educational psychology tests with regard to Bloom's Taxonomy.* Paper presented at the annual meeting of the Georgia Educational Research Association, Atlanta, GA.

Clevenstine, R. F. (1987). A classification of the ISIS program using Bloom's cognitive taxonomy. *Journal of Research in Science Teaching, 24,* 699-712.

Cox, R. C. (1965). Item selection techniques and evaluation of instructional objectives. *Journal of Educational Measurement, 2,* 181-185.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row, Publishers.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers knowledge: Performance assessment a) and post-structuralist epistemology, b) from post-structuralist perspective, c) and post-structuralism, d) none of the above. *Educational Researcher, 23,* 11-18.

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32,* 179-186.

Gall, M. D. (1970). The use of questions in teaching. *Review of Educational Research, 40,* 707-721.

Gronlund, N. E. (1993). *How to make achievement tests and assessments* (5th ed.). Boston, MA: Allyn and Bacon.

Lapointe, A. E. (1984, April). *Danger: Work on higher levels.* Paper presented at the annual meeting of the National Council on Measurement in Education. (ERIC Document Reproduction Services No. ED 243 944).

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory Into Practice, 32,* 131-137.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 22,* 5-12.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know.* Boston, MA: Allyn and Bacon.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher and school influences. *American Educational Research Journal, 30,* 523-553.

Rinchuse, D. L., & Zulio, T. (1986). The cognitive level demands of a dental school's predoctoral, didactic examinations. *Journal of Dental Education, 30,* 167-171.

Smith, C. W. (1984, August). *Verbal behavior and classroom practice.* Paper presented at the International Conference on Thinking, The Graduate School of Education, Harvard University. (ERIC Document Reproduction Services No: ED 260 077).

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 51-59

The Selection of Female Secondary School Assistant Principals and Transformational Leadership

Ann Hassenpflug
The University of Memphis

Research on school reform suggests that future school leaders need broad vision and knowledge of research on educational issues. Some researchers have suggested that hiring females as administrators is the way to bring about change and improvement in instruction in schools. This qualitative study based on interviews with recently-hired female secondary school assistant principals suggests that the females being hired into entry-level administrative positions are primarily interested in advancing their own careers, not in bringing about change. They were granted admittance to administration as a reward for demonstrating loyalty to the current administrative regime and the status quo.

The need to improve schools by selecting transformational leaders to replace the current wave of retiring administrators has been a constant theme in recent educational writing (Spady, 1985). Nevertheless, despite the need for reform in recruitment and selection of administrators identified in reports by the National Commission on Excellence in Educational Administration, the Danforth Foundation, the National Policy Board for Educational Administration, the Southern Regional Consortium of Colleges of Education, and the National Commission for the Principalship (Duke, 1992), research on the selection of new administrators has been extremely limited (Schmitt & Schectman, 1990).

Although the assistant principalship has been recognized as the key position providing access to the principalship (Haven, Adkinson, & Bagley, 1980; Marshall, 1985a), researchers have given little attention to how assistant principals are selected and how the specific process used affects the type of entry-level administrator chosen. Gips' 1988 study in which she conducted interviews with principals of twenty nonurban Ohio high schools about the recruitment and selection of their assistant principals is virtually the only study of assistant principal selection.

Gips found that principals do not generally play a significant role in the selection of their assistant principals. Superintendents played a greater role in the selection. She also found that there were higher standards used in the selection of female assistant principals.

Although Edson (1988) has studied the female administrative aspirant, little research exists on actual female entry-level administrators. Shakeshaft (1989) noted that the research on men in educational administration is insufficient for understanding women administrators. Studies by Hemphill, Griffiths, and Frederiksen (1962) and by Frasher and Frasher (1979) suggested that since women tend to outperform men in school administration, women should be preferred for administrative positions. However, Johnson and Douglas (1985) discovered that women were not promoted because it was believed that it was more efficient to hire a man. Similarly Ortiz (1982) concluded that school districts need to establish legitimate reasons for placing women in administrative positions.

More research on the selection of secondary school assistant principals is needed to determine who is selected as well as how they are selected. Such research is important in helping to ascertain the availability of potential transformational leaders.

Description of the Study

This study of the selection of recently-hired female assistant secondary school principals was undertaken to determine who was being selected as entry-level administrators and how they were being selected. Such research can be used to explore and predict how the newly hired female administrators might or might not fulfill the need for transformational leadership in secondary schools.

Participants

Twelve newly-hired female secondary school assistant principals were contacted for interviews. They were selected either from a list of new members of the

---

Ann Hassenpflug is an Assistant Professor in the Department of Educational Leadership, College of Education, at the University of Memphis. She may be contacted at Department of Educational Leadership, College of Education, University of Memphis, TN 38152.

New Jersey Principals and Supervisors Association or from schools in New Jersey and suburban Philadelphia in which the researcher's college placed student teachers. Two of the assistant principals were African-American; ten were Caucasian. The researcher's only contact with each assistant principal was the interview.

The 12 female assistant principals worked in 12 districts located in ten counties in New Jersey and Pennsylvania. All the women in this study became secondary school assistant principals in 1991-92 or 1992-93.

The student population of the districts represented a variety of socioeconomic and ethnic backgrounds. Urban, rural, and suburban schools were represented. The median student population of the districts was 3,700. Eleven of the 12 women worked in districts that ranged between 1,300 and 7,500 students. The twelfth worked in a district with over 11,000 students. The size of the secondary schools they worked in ranged from 320 to 1,600 students with the median size being 880.

Ten of the women worked in K-12 districts and two in K-8 districts. Seven were high school assistant principals while five worked in middle or junior high schools. The K-12 districts had either one or two high schools and one or two middle schools while the K-8 districts had only one middle school.

Four of the 12 new female assistant principals were the only assistant principals in their schools. These four worked in middle or junior high schools. Of the remaining eight women, only one worked in a school that had a second female assistant principal. In that instance, the middle school had four assistant principals and a female principal.

Data Collection

The interviews were loosely structured, an appropriate methodology for qualitative research (Marshall & Rossman, 1989). Each assistant principal was asked to narrate the story of her selection in her own words. She could focus on what was important and interesting to her. As each story unfolded, the role of the sole researcher was to take notes and to ask for elaboration and clarification, especially for the chronology of events (Spradley, 1979). At the conclusion of the interview, when the respondent felt more comfortable with the interviewer, she was asked for information about her career which she had not already provided in the course of the interview.

The interviews were conducted between November 1992 and March 1993 at each assistant principal's school at a time selected by the assistant principal. The length of the interview varied according to the complexity of the participant's experiences; however, most lasted 1¼ hours.

Data Analysis

The researcher's notes from each interview were typed immediately following the interview. A preliminary coding of responses and analysis according to emergent themes occurred at that time. As additional interviews were completed, patterns in thematic categories were continually revised and analyzed (Glaser & Strauss, 1967).

At the completion of all the interviews, the responses were reexamined and categorized according to the thematic patterns that had emerged from the content of the interviews (Marshall & Rossman, 1989). The results of the data collection are reported as a descriptive narrative presented according to thematic patterns in the responses as perceived by the researcher (Taylor & Bogdan, 1984).

The Route to the Secondary School Assistant Principalship

District Service

Research on female administrators (Ortiz, 1982; Prolman, 1982; Gross & Trask, 1976) has suggested that before women enter administration they have longer presocialization periods as teachers than men do. Haven, Adkinson, and Bagley (1980) found that the average woman principal worked for fifteen years as a teacher while a male principal worked only five.

At the time of the interviews the median age for the 12 female assistant principals in this study was 44 years. The median number of years for which they were assigned solely teaching responsibilities was 15.5. For the eight women who were hired from within the district where they had taught for their entire careers, the median years of service to the district was 19.5. Similar to Ortiz' findings (1982), these women had rarely moved directly from teaching to the assistant principalship. Instead, they had had supervisory or administrative positions which may have included some teaching.

Although the most typical career path to the assistant principal for these women was through years of service to their district, four of the women were outsiders. Three of the out-of-district women had been informed either directly or indirectly by their previous superintendents (two females and one male) that there was no future for them in the district's administrative hierarchy. Consequently they had applied for an administrative position in nearby districts. The fourth outsider (one of two minority women in the study) had come from a government position and had been actively recruited by the district because it wanted to hire a minority.

Administrative Certification

All the female assistant principals had principal certification at the time they were selected. Most had been certified for five to ten years before they began to apply for administrative positions. All had at least a master's degree, but the majority indicated no interest in acquiring a doctorate. They regarded having a doctorate as a disadvantage in competing for an entry-level administrative position like an assistant principal because building principals in small districts typically did not have doctorates and would be unlikely to hire a subordinate who did. The two pursuing doctorates were doing so through a national noncampus university and regarded the doctorate as helpful for future promotions.

To obtain the required master's degree and principal certification, most of the women had attended part-time graduate courses at a local state college or taken extension courses offered by a state college at a local school site while teaching full-time. Although all had to take some educational administration courses to obtain administrative certification, only two of the women had master's degrees in educational administration. Other majors for the master's degree included reading, special education, elementary education, general education, foreign language, student personnel services, urban education, and speech and language pathology.

Some of the women expressed pleasure in the camaraderie they experienced in their graduate educational administration classes and indicated that part of the motivation to take the courses was their enjoyment of going to school and having goals to which their educational effort could be directed. Others spoke of dismay with their experience in graduate courses that they felt did not prepare them for the realities of administration or for the issues that female administrative aspirants and administrators would face. All complained that their professors were male and did not show interest in their concerns. None indicated that a professor had any involvement in assisting the women in gaining an administrative position. Male professors at a research university actively discouraged one of the women from taking administration courses and pursuing an administrative career.

Instructional Certification

For their undergraduate education all but one of the women had attended colleges in their home states. The women had not ventured very far from home in pursuit of their first teaching position. All had grown up in the tri-state area of New Jersey, Pennsylvania, or New York. Typically their teaching careers began in a district that bordered on or was only a few miles from the one where they had gone to high school. For those few for whom this was not the case, the location of a spouse's job was the usual explanation for their move to the area in which they currently worked.

Only four of the 12 secondary assistant principals had teaching certification for grades 7-12. Those four were certified in academic subjects, such as English or mathematics. Elementary teaching experience and K-8 teaching certification provided entry into secondary administration as readily as secondary school teaching experience and 7-12 teaching certification. However, women with K-12 teaching certifications in physical education, special education, or vocational education were especially favored in the selection process.

The middle school assistant principalship emerged as a very competitive administrative position for women since women from both elementary and secondary schools are eligible for it. Elementary teachers were seen by all the women as being preferred over secondary teachers for middle school positions. The women thought selectors believed that elementary teachers came from a more caring, child-centered background than did a high school subject specialist.

One female assistant principal's experience with the selection process suggests that evidence of instructional leadership may actually be detrimental to an administrative aspirant. Prior to being hired as a high school assistant principal for curriculum and instruction, she was rejected from a middle school curriculum coordinator/assistant principal position in the same district. She believed her skill in instructional leadership as demonstrated in her work as a high school department chair had actually caused her to be rejected for the position. The male middle school principal hired a former male high school athletic director whose lack of expertise in curriculum guaranteed that he would do exactly what the principal told him to do. She believed the principal had wanted a yes man rather than a person who would take initiative. Shakeshaft (1989) suggested that men rate less threatening women higher than those seen as more competent.

Evidence of instructional leadership appeared to be of less importance in the selection process than was evidence of being able to handle students. Most of the female assistant principals, especially those in middle/junior high school, considered discipline to be their main task. Rosser (1980) noted that it was because discipline was a major component of the assistant principalship and because men were perceived to be better at maintaining discipline that women were less likely to be hired for this position.

Perhaps it is not surprising then that majors in physical education, special education, or elementary education were most typical for the female assistant principals in this study. Teachers who had demonstrated successful disciplinary experience handling students with physical and emotional misbehavior were favored in the selection process. Also the tasks associated with these teaching areas are (or are perceived to be) more similar to administrative tasks than are tasks of more specialized academic teaching positions which may focus as much or more on the subject matter than on student conduct.

Getting Attention of Superiors

One former physical education teacher suggested that teachers in her subject have an additional advantage in the selection process because they were perceived as "team players who could speak the language and play the games" of the male administrative hierarchy. They had more pre-selection access to the male middle and high school administrators who like themselves had been coaches.

Coaching experience may have reassured male selectors who might have felt some community pressure to hire a female assistant principal but wanted one that was as close to the male assistant principal model as possible to maintain their own comfort level. Their visibility as coaches may also have made these women stand out in the selection process since their behavior outside the classroom was already known to the community and to the selectors.

All the women, not just the former coaches, mentioned being involved in activities that gave them out-of-the-classroom experiences prior to applying for the assistant principalship. Ortiz and Marshall (1988) refer to this behavior as "Getting the Attention of Superiors." Although several districts provided opportunities for administrative aspirants to gain administrative experience via internships, summer school and night school principalships, and short-term acting assistant principalships and principalships, only one district offered a formal academy for training in-district aspirants, and that program had lasted only two years.

All of these experiences required additional time commitments from the administrative aspirants with little or no additional pay. Most of the women mentioned being asked to accept responsibilities they did not really want or were not interested in in order to signal their interest in administration to their superiors.

Entry into administration was seen as a personal challenge the women had set for themselves and met. Only one woman stated that she had entered administration to change or improve educational practices. Instead most had simply reached a point in their professional and personal lives and wanted to do something different. The obvious next step to them had been the assistant principalship. They aspired to that position because it was there. The position merely represented a step on a career ladder even though the women had not yet determined what would be the following step. The assistant principalship granted admittance to a hierarchy of administrative career opportunities that might occur if the women continued to display behavior acceptable to those above them.

How The Women Explained Their Selection as Assistant Principals

Loyalty

The issue of loyalty to the district and to the superintendent was a theme that recurred throughout the interviews. The women saw loyalty to the current regime as a necessary prerequisite for gaining an administrative position. Even when some of them felt antipathy toward central office administrators, they still felt it was extremely important not to let anyone know that.

As Miklos (1988) contended, loyalty to the person in the superintendency is a critical factor in the selection of inside candidates as assistant principals. Questioning or disagreeing with district authority figures was thought by the women to prevent selection for the assistant principalship and further promotion. Even when the women were sure that they had been rejected in favor of a less capable male for an earlier administrative position, they did not ask questions or file a grievance, because they feared such actions would harm their future chances at an administrative position in their district.

As a reward for their loyalty to the district and the administrative regime, the women did expect to be rewarded with a promotion eventually. They were willing to be patient as long as they felt they were in the district's game plan for administrative positions. At some point after signaling their interest in administration, a superior had indicated to them that their turn would come.

As several women noted, however, loyalty was not enough to ensure a female administrative aspirant an administrative position if, as one woman noted, "a veteran male superintendent had surrounded himself with an old boys' group and liked it that way." Women appeared to have a better chance at being selected as assistant principals in such a fiefdom after the despot had retired and a younger non-insider superintendent was hired as a replacement. This new person was interested in hiring his own team and was not bound by district political ties.

Similarly, an upheaval in a high school or middle school regime could also lead to increased likelihood that a woman might be hired, especially if the new principal wanted to send a clear signal to the staff that times had changed. According to one female assistant principal, hiring a female was meant to be seen by the faculty as symbolizing that a change in regime had occurred. However, according to her this change referred more to operational issues than to instructional or curricular issues.

District Interpersonal Relationships

Districts in which women held influential central office administrative positions, including but not limited to superintendent, appeared to be receptive to assistant principals being female. One female assistant principal suggested that the reason her upper middle class district had a tradition of hiring female administrators at all levels (except at the superintendency) was because the board of education knew that women could get the job done.

Two middle school assistant principals who did work in districts with recently hired female superintendents did not regard themselves as receiving any favoritism from these women, however. Their experiences with the female superintendents suggested that those superintendents actually tended to overlook the concerns of subordinate female administrators and expect automatic loyalty from them without providing any support in return.

Mentorship of any sort was rarely available to most of these women. Unlike Shakeshaft (1989), however, the women did not see that lack of mentorship was a barrier to administration. When a female assistant principal occasionally did speak of having a mentor in the district, that person was invariably a male and usually an older principal or superintendent who had originally hired the woman as a teacher.

Political connections within the district were regarded as more important than sponsorship. Working for administrators who later were in a position to influence directly or indirectly who was hired as an assistant principal was extremely important. Examples of such administrators included male assistant principals who later became middle school or high school principals, a male elementary principal who served temporarily as an acting middle school principal and consequently had close ties with the male middle school principal who succeeded him and had to hire an assistant principal as one of his first duties in his new district, and a male assistant principal who served on a selection committee for a new assistant principal.

The women also mentioned having informal support systems that might include other administrative aspirants who were either male or female, older administrators who were nearing retirement, and spouses who were either administrators or aspiring administrators. Some of these same women who had informal support from uninfluential individuals for their administrative aspirations noted that they had encountered obstacles and both overt and covert hostility from administrators of both sexes in their districts even before they developed an interest in seeking an administrative career. The high school academic subject teachers believed their ability and effort were regarded as threatening to those in power even when they were teachers.

Gender Issues

The female assistant principals in this study reported few instances that they labeled as examples of discrimination against females even though they described situations that appeared to the researcher to be discriminatory. They seldom acknowledged encountering sexist behaviors or attitudes even though their descriptions of situations suggested that such a label would have been accurate. This finding is contrary to Edson's (1988) study of female administrative aspirants who claimed to have encountered tokenism, game playing, and biased questioning in the selection process.

The female assistant principals' narrow focus on their personal career advancement appeared to have made them wary of being involved in any controversial women's issues that might have a negative effect on their careers. According to one woman, she was hired because "she keeps things on an even keel" rather than protesting slights or criticizing male administrators for their behaviors or comments. One middle school assistant principal commented that she would have refused the other vacant assistant principalship (eventually filled by a male) in her district because of the reputation of the school's male principal as a sexist. She explained that her decision was not meant as a political statement but rather an expression of a desire to avoid confrontations and unpleasantness.

Although a few women in districts which previously had few or no female administrators thought that being female had been an advantageous factor for them in the selection process as indicated by one assistant principal's comment that she had heard rumors that the job she eventually got "was slated for a woman," the female assistant principals did not regard affirmative action as a major reason for their selection. The women believed that the main reason they were hired was that their experience and skills fit the current needs of the district.

The women did not view themselves as tokens. Only one woman acknowledged that she was recruited and hired specifically because of who she was. This minority female assistant principal knew at the time of the selection process that the district was specifically looking for a minority assistant principal although not necessarily a female.

District Financial Issues

District financial constraints emerged as a factor in some of the decisions to select a female assistant principal although few of the women pointed out the relationship. For instance, several women continued to perform their previous duties as supervisor of one or more departments or subject areas for the school or the district in addition to performing the duties of an assistant principal.

The female assistant principals' comments about the job titles and salaries of the other finalists for the position revealed that a female might cost the district less because her placement on the administrative salary schedule would be lower than male finalists who had higher salaries as teachers due to more seniority or extra-curricular duties.

According to one middle school assistant principal, the board of education had determined the top salary for the vacant assistant principalship prior to the selection process. She noted that she was the only one of the four in district finalists who would not have to take a lower salary to move from teaching to administration.

The Selection Process as Portrayed by the Assistant Principals

The women had little experience with administrative selection processes prior to the one for their current assistant principalship. Due to their limited experience they tended to assume that the steps used in the selection process used in their district was standard for all districts. However, the processes they described to this researcher demonstrated that the processes varied significantly between districts. Although all the processes relied heavily on personal interviews, there was no standardization in regard to the quantity of interviews, the identity of interviewers, or the sequencing of interviews when the selection process involved a series of interviews. Additionally, the timeline for completion of the steps in the process could be limited to a week or extend over several weeks or months.

Recommendations for Changes in the Process

Although all the women claimed that the selection processes used by their current district in hiring an assistant principal were fair, they insisted that the processes were in need of revision. They questioned the value of some of the steps and procedures they encountered during that process. Preliminary screening interviews conducted by committees were repeatedly described by the women as "an interrogation" or "an inquisition." One assistant high school principal described the interview committee as using "an aggressive line of questioning."

The women also complained about being made to feel like an outsider during committee interviews. One future assistant high school principal was required to sit in a student desk facing a semicircle of committee members seated in armchairs. Another woman noted that the interview committee had a tray of food available which she was never invited to share.

The women much preferred having an interview with just the principal or the principal and superintendent. They felt most comfortable during such interviews when they began to "feel like conversation." Meeting with the principal was the most important step in the process for the women. Getting a feel for who the principal was and not just trying to impress him or her with who they were was important to all of the women. They wanted to make sure they "could work with the principal" and that "there was the right chemistry."

Repeatedly the women spoke of the frustration over the lack of communication from the central office about the selection process. The unpredictability in the manner in which the women were informed about their selection was a source of frustration. For instance, one middle school assistant principal was unexpectedly summoned into the superintendent's office during her unscheduled visit to the central office to work on the budget for a subject she supervised. A high school assistant principal was awakened late at night by a phone call from the principal to inform her of her selection by the school board.

All the female assistant principals, regardless of the selection process used to hire them, commented on the "political nature of the process." For them politics meant the micropolitics of the school board, the superintendent, other district administrators, the principal, and school faculties. Their knowledge about the role of these specific players was limited, however. Most thought it was not to their advantage to be curious about how their selection had occurred

Secrecy

The secrecy cloaking the selection process emerged as a persistent theme as the women talked about their experiences. They felt they could not and should not ask questions about any steps in the process other than perhaps the timeline for making the decision. Questions about how many applicants there were, how many candidates there were, how many and who the finalists were, and who would be present at the interviews were regarded as inappropriate to ask outright. However, if a candidate knew a secretary, other candidates, or some of the interviewers, then these sources could be asked for information behind the scenes.

Even when the information was given to the candidates at the interview, it could change or turn out to be inaccurate. For instance, they might be called to additional interviews, the timeline for the decision might be significantly extended, or the school board members might become more involved personally or as a group than was originally indicated.

Although the women may have discussed parts of the selection process with other candidates or with their own principals during or after the process, the interview for this study was the first time any of them had talked about the selection process in detail and reflected on it. They claimed that they found the interview to be a therapeutic experience, especially when they had pent-up feelings about unfair practices, district politics, and conflicts with other administrators that they had encountered during the selection process.

As a further example of the secrecy surrounding the selection processes, few women knew who actually made the decision to hire them. They weren't sure if it was the principal, the superintendent, the superintendent and the principal jointly, or the school board. One high school assistant principal commented that she felt the principal didn't want her but since she arrived, she has won him over.

The women suspected that despite increasingly formalized selection processes, hiring decisions are just as often made before their interviews as after. This view was not limited to the insider candidates. One assistant principal who had been an outsider candidate was later told by her principal that the principal had made the decision to hire her after their first phone conversation. Ironically, this assistant principal had to go through the most outwardly complex selection process of any of the women. She was interviewed by four separate committees composed of district and building administrators, board members, students, community members, faculty, and parents.

Insider Candidate Advantage

Regardless of whether they had been insider or outsider candidates, the assistant principals believed an insider candidate had a better chance. The only time they thought an outside female had a significant chance was if all the internal candidates had been ruled out prior to the interviews or if the district were overtly or covertly looking for a minority or a female.

Although the women regarded interviews as anxiety-producing experiences, they assumed they were a necessary step in the selection process. They did, however, resent having to sit through interviews when they already knew the district had selected someone else. While some saw an advantage in going to interviews that led to rejection because influential people on the interview committees would be more knowledgeable about their abilities when they applied for future administrative openings, most did not see any merit in just going to interviews to gain experience in interviewing.

The only preparation they did for the interviews was to read district policies, think about how they might answer questions on certain topics they expected to be asked, and carefully select professional attire. As one assistant principal said, "If being myself isn't what they wanted, then the job wasn't for me." Only one assistant principal, an outsider candidate, took any additional materials about herself to the interview to share with the interviewers. She was the only one who sent a thank you note to the interviewers.

Implications for Transformational Leadership

The limited number of participants in this study as well as the limited geographic area create some limitations in suggesting implications regarding the relationship between hiring entry-level female secondary school administrators and school transformation. Further research involving larger numbers of participants, a larger geographic area, and other methodologies need to be used to examine the issues that emerged in this study.

A major concern for school transformation that emerges from this study is whether an individual who is selected as an assistant principal because she fits well with the current regime and is not perceived as a boat-rocker is able to raise potentially controversial questions and demonstrate the visionary leadership necessary to bring about school change. The view of administration as being about efficiency and control (Ortiz & Marshall, 1988) appears to remain firmly entrenched in these districts. Little appears to have changed in administrative selection processes since Baltzell and Dentler (1983) concluded from their study of thirty principal appointments in ten districts that the primary criterion for selection as a building administrator was the perceived "fit" between the candidate and the current image and culture of the school or the district.

Although the female assistant principals spoke disdainfully of other administrators they regarded as supporters of the status quo, they perceived their own choice of silence and avoidance of risks as acts of self-protection rather than as submissive behavior. Getting along with men, having hobbies and extracurricular involvements that were similar to men's, and being seen as one female assistant principal said, "as a clone of the previous male assistant principal" were regarded as very desirable traits for women aspiring to the position. Marshall (1985b) suggested that strong undercurrents in both the formal and informal structures of schools force women to learn to "pass" as men. Consequently, they are forced to gravitate toward male postures, male behaviors, and even male clothing to gain admittance to the field.

Shakeshaft's (1989) recommendation that hiring more female administrators will increase the likelihood of improving school is too simplistic. Not only are the females being selected the ones who most resemble the reigning male administrators in regard to behaviors and interests, but also most of these female assistant principals indicated little interest in instructional issues or in innovation.

The few women who did show interest in dealing with substantive educational issues like curricular reform or program development typically had either high school teaching experience in an academic subject or a strong background in elementary gifted education. This small group of assistant principals included women whose job responsibilities as assistant principals consisted mainly of teacher supervision and curriculum development, while a male, either another assistant principal or the school principal, handled operational tasks and student discipline. Although the women in such instructional leadership positions regarded them as more desirable and important than the typical male assistant principalship, which one female called "a sweep up position," they also acknowledged that such a position was probably a dead-end position for them if they chose to remain in their current district

It is also questionable whether most of these women's limited perspectives and limited knowledge base regarding educational reform could evolve into the vision considered necessary for leadership. Their preparation in educational administration as well as their own professional development efforts were minimal. Perhaps as a result of the lack of a master's degree in educational administration, most of the women's comments demonstrated limited knowledge about current issues in educational administration related to improving schools as organizations and to providing learning opportunities to students. Few read educational journals, attended conferences, or had regular communication with administrators in other districts. The only professional organization in which some of them had taken an active role as a teacher or an administrator was a district union.

Although women might be classified as nontraditional candidates for administrative positions, merely hiring women does not necessarily mean reform is going to occur. Women selected for entry-level administrative positions to meet district public relations or financial needs display a relatively narrow focus and limited interests. They are more interested in advancing their own careers and having new personal growth experiences than in improving schools. They appear to have been selected because they offer little threat to a status quo from which they have benefitted and hope to continue to benefit.

This study suggests that current selection processes for entry-level administrative positions need to be reexamined and redesigned if candidates with the potential for visionary and instructional leadership are to be selected. Systemic reform in the form of changing district practices that support selection of nontraditional candidates for leadership positions needs to be adopted by school districts desiring candidates who demonstrate instructional leadership rather than conformity to district standards and politics.

References

Baltzell, D., & Dentler, R. (1983). *Selecting American school principals: A sourcebook for educators.* Washington, DC: National Institute of Education.

Duke, D. (1992). The rhetoric and the reality of reform in educational administration. *Phi Delta Kappan, 73*(10), 764-70.

Edson, S. (1988). *Pushing the limits: The female administrative aspirant.* Albany, NY: State University of New York Press.

Frasher, J., & Frasher, R. (1979). Educational administration: A feminine profession. *Educational Administration Quarterly, 15*(2), 1-13.

Gips, C. J. (1988). Selection of secondary school assistant principals: Selected case studies. *The High School Journal, 71*(4), 167-177.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Gross, N., & Trask, A. (1976). *The sex factor and the management of schools.* New York: Wiley and Sons.

Haven, E., Adkinson, P., & Bagley, M. (1980). *Women in educational administration: The principalship.* Annadale, VA: JWK International Corp.

Hemphill, J., Griffiths, D., & Frederiksen, N. (1962). *Administrative performance and personality.* New York: Teachers College Press.

Johnson, M., & Douglas, J. (1985). Assessment centers: What impact have they had on career opportunities for women? *NASSP Bulletin, 69,* 110.

Marshall, C. (1985a). Professional shock: The enculturation of the assistant principal. *Education and Urban Society, 18*(1), 28-57.

Marshall, C. (1985b). The stigmatized woman: The professional woman in a male sex-type career. *The Journal of Educational Administration, 33*(2), 131-152.

Marshall, C., & Rossman, G. (1989). *Designing qualitative research.* Newbury Park, CA: Sage.

Miklos, E. (1988). Administrator selection, career patterns, succession, and socialization. In N. Boyan (Ed.), *Handbook of research on educational administration* (pp. 53-76). White Plains, NY: Longman.

Ortiz, F. (1982). *Career patterns in education: Women, men and minorities in public school administration.* New York: Praeger.

Ortiz, F., & Marshall, C. (1988). Women in educational administration. In N. Boyan (Ed.), *Handbook of research on educational administration* (pp. 123-141). White Plains, NY: Longman.

Prolman, S. (1982). Gender, career paths, and administrative perceptions. *Administrator's Notebook, 30*(5), 1-4.

Rosser, P. (1980). Women fight "Old Boys" for school administrator jobs. *Learning, 8*(7), 31-34.

Schmitt, N., & Schechtman, S. (1990). The selection of school administrators. *Journal of Personnel Evaluation in Education, 31,* 231-238.

Shakeshaft, C. (1989). *Women in educational leadership.* Newbury Park, CA: Sage.

Spady, W. G. (1985). The vice-principal as an agent of instructional reform. *Education and Urban Society, 18*(1), 107-120.

Spradley, J. S. (1979). *The ethnographic interview.* New York: Holt, Rinehart, and Winston.

Taylor, S. J., & Bogdan, R. (1984). *Introduction to qualitative research: The search for meanings* (2nd ed.). New York: John Wiley.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 61-66

A Survey of Accelerated Master of Teaching Program Graduates at The University of Memphis

Tiffany L. Bailey and Linda Bol
The University of Memphis

A survey was developed to provide follow-up information about graduates of the accelerated Master of Teaching (M.A.T.) program at The University of Memphis. Questionnaires were mailed to every graduate of the program since its inception, with forty-one percent (N=66) of the graduates responding. Results showed that most graduates tended to be hired immediately after graduation and remain in teaching for five years or more. Though some graduates pursued additional degrees or careers, they were in the minority. Overall, graduates perceived that the program prepared them well for teaching and also identified many program strengths among which were the long internship, university professors, and challenging yet practical content. Some weaknesses, most notably lack of adequate preparation in classroom management skills, were identified.

As the United States has become increasingly aware of the status of education within the country and in comparison to other countries, a surge in reform of various types has marked the last three decades in an unprecedented manner. Since the origination of this renewed consciousness in education, a national focus of reform has consistently been in teacher education. The ensuing trend was not only an impressive increase in the number of universities offering fifth-year masters programs for aspiring secondary teachers, but simultaneously, an explosion of new master's degree programs totaling over 100 master's titles nationally by 1989 (Osguthorpe & Wong, 1991).

The Masters of Art in Teaching (M.A.T.) model for teacher education emerged during the mid-seventies as a way to enhance the quality and effectiveness of teacher education programs. This typically demanding and fast-paced program was designed with the specific goals of recruiting and maintaining liberal arts majors in the field of education by establishing higher standards of acceptance to and proficiency within this teacher education program (Vollmer, 1986). The independent nature of universities in developing their own programs without a national or statewide standard, however, has led to a large variation on the original concept of an M.A.T. program with many functioning and constantly evolving models at universities across the country (Osguthorpe & Wong, 1991).

Although the existence of Master of Arts in Teaching programs can be traced back to the 1930's at universities such as Harvard and Northwestern, the last three decades have seen a revitalization of the M.A.T. model (Kauss, 1988). In their study of 664 United States institutions with graduate programs in education, Osguthorpe and Wong (1991) found that the M.A.T. degree has been the fourth most popular degree title since at least 1979, while still showing positive growth patterns by 1989. One of the side effects of the degree's popularity, however, is the increasing diversity of curricular demands and other requirements among universities and even within universities as the programs become more tailored to the individual needs of the institutions.

Whereas the literature addressing the success or effectiveness of these programs is sparse, the available research relevant to the evaluation of these programs is positive. In a case study comparing an M.A.T. program with a certification only program, Arch (1989) found that the M.A.T. degree candidates received higher overall classroom evaluations by their public school evaluators during the practicum than did first-year teachers from undergraduate certification programs at the secondary level. Eichelberger and Bean (1990) reported that the more progressive and innovative institutions in education that phased-out their entire undergraduate programs in education for the more advanced level of licensure in a fifth-year program acknowledge the stronger content and

---

Tiffany L. Bailey is a graduate of the Accelerated Master of Teaching Program at The University of Memphis. She is currently a teacher at Overton High School in Memphis. Linda Bol is an Assistant Professor of Educational Psychology and Research at The University of Memphis. Correspondence regarding the article may be sent to the authors at the Department of Counseling, Educational Psychology and Research, College of Education, The University of Memphis, Memphis, TN 38152. Email: Bol.linda@coe.memphis.edu.

professional knowledge of these degree recipients than those from traditional undergraduate certification programs. Other studies support these positive evaluation findings (Freeman, 1985; Kauss, 1988; Niebrand, 1992, Soldwedel, 1984).

The most comprehensive study on the profiles of M.A.T. students degree recipients was sponsored by the Ford Foundation. In this study, Coley and Thorpe (1985) identified many trends among the students at the universities offering an M.A.T. program including race, sex, age, reasons for pursuing the M.A.T. degree, academic achievement, educational plans, and career plans. They found that the majority of M.A.T. degree recipients were white females from middle to upper socioeconomic classes who were in their mid-twenties. The graduates cited the desirability of receiving certification along with a salary-boosting master's degree as the primary reasons for applying to the M.A.T. program. The relatively short period of time required for attaining the degree was also cited. Those entering the program tended to be at least B+ students in their undergraduate work, and many went on to earn other degrees. Although only one-third of the graduates were employed in a teaching position just five years after earning the degree, about another third remained in education in some capacity. Though the sample of institutions was not random, these results shed some light on the characteristics of M.A.T. graduates and the program's success at training educators, at least for this sample of schools.

Although the literature provides some information about the success of M.A.T. programs, the characteristics of M.A.T. degree recipients, reasons for enrolling in the program, and employment trends after graduation, most studies are somewhat dated and do not address the evaluation of the program from the perspective of program graduates. Clearly, more current information about program graduates and their evaluation of the program is needed. The present study provides this information for graduates of the accelerated M.A.T. program at The University of Memphis. This is the only academic institution in the tri-state area (Tennessee, Arkansas, Mississippi) which offers an accelerated licensure plus masters program, and though the results may not generalize to all M.A.T. programs nationwide, they provide an important case study of graduates from this program.

The University of Memphis' version of a fifth-year masters program in education has been evaluated from the perspective of program mentors and current students. For example, one study focused on the first years of the program's inception (Butler, 1986). The evaluation of the program was from a mentor perspective attained through surveys and interviews with the university and classroom mentors. Other research projects included surveys of current students in all teacher education programs in the college, including M.A.T. students (e.g., Dietrich, 1995; Etheridge, 1995; Etheridge, Butler, Etheridge, & James, 1988). However, there has been no follow-up research on graduates of the accelerated M.A.T. program that asks why the M.A.T. graduate students decided to enroll in the program, what they thought about the program, and what they did with their masters degree.

The purpose of this study was to obtain follow-up information about the graduates of the program and their perceptions of the accelerated M.A.T. program at The University of Memphis. More specifically, the research questions addressed included:

1. Why did students choose a) The University of Memphis' M.A.T. program and b) the accelerated program over the self-paced program?
2. How soon after completing the accelerated M.A.T. program did these graduates find employment? What kinds of employment did these graduates obtain?
3. What are the future employment and educational plans of these graduates?
4. Did accelerated M.A.T. graduates perceive that they were well-prepared for teaching?
5. What do graduates of the program perceive as the strengths and weaknesses of the accelerated M.A.T. program?

Method

Participants

The questionnaire was mailed to all graduates of the accelerated M.A.T. program at the University of Memphis since its inception in 1985 until 1993. Of the 185 graduates of the program, questionnaires were mailed to the 161 graduates with available addresses. A total of 66 graduates completed the surveys for a response rate of 41%. Respondents included graduates from every graduating class.

Questionnaire

The questionnaire items were developed to address the research questions and included both open-ended and close-ended items. Demographic items asked respondents to identify their year of graduation, their age at graduation, their sex, and their ethnicity. The close-ended items consisted of Likert-type ratings, checklists, and forced-choice items. For open-ended items, a qualitative content analysis was employed to identify patterns of responses that were used to develop categories of responses. Responses were then coded based on these categories, and frequencies and percentages were tabulated by category.

Procedure

Surveys were distributed by mail with a cover letter and an addressed/stamped return envelope. Each survey was coded to maintain confidentiality of respondents while allowing a follow-up mailing to non-respondents. After three weeks had elapsed since mailing the questionnaire, those subjects who had not returned the survey received a second letter and another copy of the survey with a return envelope.

Results

Demographic Characteristics of Respondents

Age at the time of graduation from this program ranged from 22 to 51 years. Although the mean age was 31.7 years (*sd* = 8.82), the median was 28.5 years, indicating that 50% of the respondents were between the ages of 22 and 28.5 years at the time of graduation.

The largest ethnic group was Caucasian, representing 81% of the respondents. Only 8% of the respondents were African American, and only 3% were Native American. While no other ethnic group was identified, 8% selected the "Other" category.

There was a higher percentage of female compared to male respondents. Women comprised 64% of the graduates while men comprised the other 36%.

In order to determine whether the demographic characteristics of respondents were similar to the demographic characteristics of program graduates, the researchers obtained application records for all candidates who graduated in 1991, 1992, and 1993. From the 74 applications, information on sex, ethnicity, and age was collected. The percentage of females versus males was identical to the percentages computed for the sample of graduates who responded to the questionnaire (64% female versus 36% male). The distribution of graduates by ethnicity was somewhat similar to the distribution observed in the sample of respondents. Most of the graduates were Caucasian (91%), while much smaller percentages identified themselves as African Americans (6%), or "Other" (3%). Because the provision of information on ethnicity was optional on the application, not all applicants supplied this information. In fact, six applicants did not identify their ethnicity. The omission of this data may help account for the difference in the distribution of ethnicity between the respondents and program graduates. The statistics for age at the time of graduation obtained from the applications were also similar to those reported for the respondents. The mean age of graduates was 29.9 (*sd* = 7.66), and the median age was 25.5. Overall, the demographic characteristics of respondents were similar to the demographic characteristics of program graduates for these years and suggest that the sample is reasonably representative of the population of students who graduated from the M.A.T. program.

Reasons for Choosing Program

The primary reason cited for the selection of The University of Memphis was the location of the campus (20% of respondents). This was an important consideration for prospective students because of proximity to their homes and to secondary schools that were potential sites for internships. The availability of an accelerated M.A.T. program at The University of Memphis was almost as important as location as the reason for enrolling at this institution. Just over 19% of graduates said that the opportunity to complete the program in such a short period of time was a major factor in choosing The University of Memphis. There is no other academic institution in the tri-state area which offers an accelerated licensure plus masters program, and many of these respondents indicated their consideration of that fact when selecting a graduate school.

When asked about why they chose the accelerated versus the self-paced program, it was not surprising that most respondents (56%) said that the duration of the program was the major reason for selecting the accelerated program. In reference to program length, some respondents noted the ability to attain employment sooner, while others believed that they had been in school long enough and would not have pursued a higher degree were it not for the limited time frame. The second most frequently cited reason (19%) was the availability of stipends to cover tuition and salary for the year of teaching/interning in a local school system. Some respondents (11%) felt that the year-long internship offered in the accelerated program was superior to the 15 week student-teaching experience required in the self-paced program because they believed that they would receive more "practical experience" by being in a school an entire year and actually having their "own classes."

Employment

As shown in Figure 1, the employment rates of graduates from the accelerated M.A.T. program were very high. Overall, 76% of respondents were hired by a school system by the beginning of the school year immediately following graduation. Of the remaining graduates, 11% were hired within 12 months of that same school year and 5% attained teaching employment two to three years after graduation. Though 8% were never employed as teachers, some respondents commented that it was because they had not applied for a teaching position. Only two respondents who had not been employed as teachers indicated that they had sought employment and not been hired.

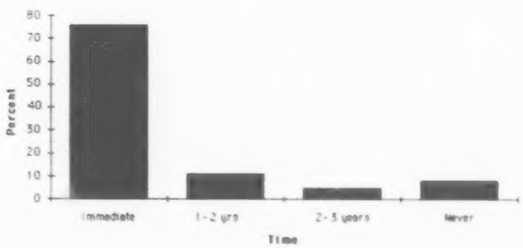


Figure 1. Time between graduation from the M.A.T. program and employment of program graduates.

The initial place of employment for these graduates tended to be in public schools (87%). As for those graduates not in public schools, 8% were employed in a private school, and 5% were hired as college instructors in their first teaching position. However, the initial employment was diverse with respect to the type of public school in which the graduates were hired. Suburban schools employed 43% of graduates with inner-city and rural schools employing 33% and 24% of the program graduates respectively. Most of these schools (90%) were located in the state of Tennessee.

Future Plans

When asked about their employment plans for the future, 71% of the graduates reported that they planned to remain in teaching, 19% planned to remain in the field of education in some capacity, and 9% planned to pursue a field other than education. Of those remaining in education, slightly more than half planned to pursue some administrative position. Almost one fourth of those remaining in education planned to do so at the university level.

None of the graduates had gone on to earn a terminal degree, although four graduates had earned an additional masters degree in another field. While almost half of the respondents reported their intentions to pursue another degree, only 10% of the graduates were currently doing so. Those graduates who were either currently pursuing a degree or were planning to pursue another degree expressed several reasons for wanting further formal education. The most frequently cited reasons included personal gratification or goal fulfillment, the aspiration to be hired as a university professor, and increased earning potential.

Because the researchers wanted to determine whether future plans differed based on the sex or ethnicity of the respondent, Chi-square analyses were computed. The results showed that there were no significant differences in the future plans of these graduates (in terms of employment or education) by sex or by ethnicity.

Graduates' Evaluation of the Program

Program Ratings. The questionnaire contained eight Likert-type items where graduates rated the program on how well it prepared them in specific teaching practices and for teaching overall. Table 1 provides the mean ratings obtained and the percentages by response category. The high mean ratings for all but one item suggest that the graduates thought that the program effectively prepared them for teaching. The only exception to this trend was the relatively low mean rating obtained for classroom management which was also identified as a major weakness of the program in open-ended comments. Based on ANOVA results, we found that ratings did not significantly differ by year of graduation or initial teaching employment (i.e., rural, suburban, or inner city school).

Strengths. Graduates were very positive about the accelerated M.A.T. program as a whole and identified several strengths of the program. The five most frequently endorsed strengths appear in Table 2. Because some responses included reference to more than one strength, the percentages do not sum to 100; the percentages were computed based on the number of graduates who responded to the item. The most frequently noted strength was the year-long internship in a local school system which benefited the interns in terms of extended practice, networking, and integrating theory with practical application. The second most frequently cited strength was the university professors. Graduates praised the professors for their content knowledge, personal concern and understanding, and effective instruction. Program content was described as practical and useful. Class camaraderie and the short duration of the program were also frequently cited strengths. This pattern of strengths was consistent across years of graduation.

Table 1
M.A.T. Program Ratings by Graduate Respondents: Number, Mean Value and Percentage of Agreement by Area

Area	n	Mean	Strongly Agree	Agree	Disagree	Strongly Disagree
Lesson planning	58	3.72	74.1	24.1	1.7	0
Testing & evaluation	59	3.54	59.3	35.6	5.1	0
Instructional methods	58	3.48	55.2	37.9	6.9	0
Content teaching	58	3.36	48.3	41.4	8.6	1.7
Academic research	58	3.25	31.0	63.8	5.2	0
Curriculum planning	59	3.23	35.6	52.5	11.9	0
Classroom management	58	2.67	25.9	25.9	37.9	10.3
Overall program	56	3.46	50.0	46.4	3.6	0

Table 2
The top five strengths and weaknesses of the accelerated MAT program identified by program graduates.

Category of Response	% of Respondents
Strengths	
1. Internship	52.38
2. University professors	38.09
3. Content	34.92
4. Class camaraderie	28.57
5. Program length	28.57
Weaknesses	
1. Lack of university support	44.82
2. Unnecessary content	43.10
3. Classroom management	18.96
4. Time consuming schedule	13.79
5. Thesis	10.34

Note. Because some respondents cited more than one strength or weakness, the percentages do not sum to 100; the percentages are based on the number of respondents.

Weaknesses. Overall, the graduates identified as many weaknesses as strengths of the program. Table 2 also shows the five most frequently cited weaknesses of the program. Again, respondents may have identified more than one weakness, and the computation of percentages was based on the number of respondents. The most frequently reported weakness was some instance of lack of university support during the program. This weakness was related both to the lack of support in obtaining employment after graduation and to the lack of communication and cooperation among the university, school administrators and classroom mentors, especially in instances where the interns perceived their placements as unsatisfactory. The inclusion of unnecessary content in the curriculum was the second most frequently identified weakness of the program. Some content was perceived as redundant, as lacking "real world" application, and as focusing on elementary school issues and examples. The remaining three weaknesses included inadequate instruction in classroom management, the time-consuming schedule, and the thesis requirement which was described as unnecessary and so time-consuming that it detracted from teaching. Again, the pattern of weaknesses that emerged did not significantly vary depending on year of graduation.

Discussion

The results suggest that the M.A.T. program was perceived as largely successful in preparing its students for teaching employment and was favorably evaluated by graduates of the program. Evidence for the program's success was based on the high employment rates of graduates, the high mean ratings obtained on the questionnaire, and the identification of several program strengths. The most frequently identified strengths were the internship experience, the university professors, and program content.

However, weaknesses of the program were also cited. A weakness that emerged both in the open-ended comments and in the ratings was classroom management. Respondents reported that not enough instruction was provided in this area and that the instruction which was provided was not appropriate for dealing with "real world" problems at the secondary level. Other weaknesses included lack of university support and unnecessary or redundant program content. Addressing these weaknesses should serve to further strengthen the M.A.T. program at this institution.

The generally positive evaluation results obtained from M.A.T. graduates at The University of Memphis add to the limited body of research that suggests that these types of teacher preparation programs are successful (e.g., Arch, 1989). Moreover, these results augment the local evaluation efforts at The University of Memphis focusing on students who are enrolled in the program rather than on program graduates (Dietrich, 1995; Etheridge, 1995).

Though the results should not be generalized to all M.A.T. programs, there were some notable parallels of the findings obtained in the present study and the Coley and Thorpe study (1985). One similarity pertains to the demographic characteristics of program graduates. The largest percentage of graduates were female Caucasians, and these demographic characteristics are similar to those reported in the Ford Foundation study (Coley & Thorpe, 1985). Though one cannot be certain whether these samples are representative of all program graduates, it does raise the question of whether the goal of increasing diversity in teacher training programs is being realized. A second similarity between the present results and those obtained by Coley and Thorpe (1985) is the reason for applying to the M.A.T. program. Obtaining certification concurrently with a Masters degree and the relatively short period of time required for obtaining the degree were cited as major reasons for selecting the program in both studies. It would appear that these reasons might be highlighted when recruiting students into programs in order to attract more qualified and diverse students.

Because of the limited amount of research conducted on graduates of M.A.T. programs, the study does contribute important follow-up information, at least for the graduates of one particular M.A.T. program. However, the study is not without limitations which illuminate directions for future evaluation research on the success of these programs. As noted previously, external validity or the ability to generalize to all graduates of the program at The University of Memphis or to graduates at other universities is suspect. A more representative sample of M.A.T. programs that includes follow-up information from a larger percentage of program graduates at each institution would provide more valid results. Moreover, comparisons of data gathered from graduates of M.A.T. programs with graduates of other teacher preparation programs would strengthen this line of research. Further follow-up studies of program graduates could help fill a gap in our knowledge about the success of this promising teacher preparation program and the reform efforts in teacher education.

References

Arch, E.C. (1989, March). *Comparison of student attainment of teaching competencies in traditional preservice and fifth-year master of arts in teaching programs*. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

Butler, E. D. (1986). *Mentor perceptions of mentoring and internships in MAT and Lyndhurst programs - cycle I.* Memphis, TN: University of Memphis, Center of Excellence in Teacher Education.

Coley, R. J., & Thorpe, M. E. (1985). *A look at the MAT model of teacher education and its graduates: lessons for today*. Princeton, NJ: Education Testing Service.

Dietrich, A. (1995, February). *Factors related to successful extended field experiences among secondary M.A.T. students*. Paper presented at the annual meeting of the Association of Teacher Educators, Detroit, MI.

Eichelberger, T. R., & Bean, R. (1990, February). *Academic and professional knowledge and skills of four-year undergraduate and fifth-year teacher certification students*. Paper presented at the Eastern Educational Research Association Conference, Clearwater, FL.

Etheridge, C. P. (1995, February). *A research design for comparative study of five teacher education programs: Undergraduate SPED & ELED, Self-paced MAT ELED & SECED, and Accelerated MAT SECED*. Paper presented at the annual meeting of the Association of Teacher Educators, Detroit, MI.

Etheridge, C. P., Butler, E. D., Etheridge, G. W., & James, T. (1988, February). *The effects of type of teacher preparation program on internships in secondary schools*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

Freeman, B. (1985). *Exploring new frontiers in teacher education: The Austin teacher program, "over a decade of pioneering" -- an exploration*. Sherman, TX: Austin College.

Kauss, T. (1988). *The master of arts in teaching*. Washington, DC: Academy for Educational Development, Inc.

Niebrand, C. (1992). Insecurity, confusion: Common complaints of the first-year teacher. *NASSP Bulletin, 76*, 546.

Osguthorpe, R. T., & Wong, M. J. (1991). *The growing confusion among master's programs in education*. Provo, UT: Bingham Young University, College of Education.

Soldwedel, B. J. (1984). *Teacher education in Florida: Models and prospects*. Jacksonville, FL: Florida Institute of Education.

Vollmer, M. L. (1986, February). *Meeting the teacher shortage head on*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Chicago, IL.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 1, 67-82

Using A Priori Versus Post-Hoc Assignment of a Concomitant Variable to Achieve Optimal Power from ANOVA, Block, and ANCOVA Designs

Yi-Cheng Wu
Taipei Municipal Teachers College

James E. McLean
The University of Alabama at Birmingham

By employing a concomitant variable, block designs and ANCOVA may increase the power over a completely randomized ANOVA. If the concomitant variable is not considered when subjects are assigned to treatments, an experiment uses a post-hoc approach; otherwise an a priori approach is used. Traditionally, a priori assignment has been considered the more powerful approach. Results of this Monte Carlo study show that the a priori approach is not generally more powerful than the post-hoc approach and ANCOVA is not always more powerful than block designs. The most powerful technique for employing a concomitant variable varies depending on the experimental conditions. However, many of the differences in power are not large enough to be of practical significance. Tables are provided to assist educational researchers in their selection of the best design for using concomitant variables.

In 1974, a study by Edgington found that 70% of all articles published in journals sponsored by the American Psychological Association made use of some form of analysis of variance (ANOVA). The Preface to the 1996 Edition of the popular statistical textbook, *Statistical Methods in Education and Psychology* (Glass & Hopkins), noted that it had increased its coverage of ANOVA techniques "because ANOVA continues to be the most widely used statistical technique in psychological and educational research" (p. xiii). While not everyone agrees that ANOVA is the uniformly most powerful approach to many analyses (e.g., Thompson, 1985, 1994), it remains a widely used analysis technique by psychological and educational researchers.

A common method used to increase the power of a simple one-way ANOVA design is to add a concomitant variable such as in block designs and ANCOVA. Whether to block or covary and how many blocks to use if a block design is chosen become crucial decisions. Wu and McLean (1993) and Wu (1994) provided an historical review of the problem, finding that some researchers favor block designs while others prefer ANCOVA. The most comprehensive studies on this topic were conducted by Feldt (1958) and Maxwell and Delaney (1984). Feldt analytically examined the problem using Apparent Imprecision as the criterion variable, while Maxwell and Delaney empirically examined the problem using the Type I error rate and power in addition to Apparent Imprecision as the criterion variable. Based on Apparent Imprecision, Feldt found the correlation coefficient between the concomitant and dependent variables should be the primary factor in choosing blocking or ANCOVA. He also suggested the optimal number of blocks to choose if blocking were used. Feldt's findings have been cited by many research articles and texts discussing this work (cf., Maxwell & Delaney, 1984; Wu, 1994; Wu & McLean, 1993, 1994b).

The recommendation to consider the correlation in choosing blocking or ANCOVA was rejected by Maxwell and Delaney (1984); "instead, the two factors that should be considered are whether scores on the concomitant variable are available for all subjects prior to assigning any subjects to treatment conditions and whether the relationship of the dependent and concomitant variable is linear" (p. 136). Since Maxwell and Delaney suggested that power might provide a different perspective from Apparent Imprecision, the number of blocks used by them, which was based on Apparent Imprecision and

---

An earlier version of this paper was the recipient of the Outstanding Paper Award at the 1994 Annual Meeting of the Mid-South Educational Research Association. Yi-Cheng Wu is an Associate Professor, Department of Arts and Crafts Education at Taipei Municipal Teachers College in Taiwan. James E. McLean is a University Research Professor and Director, Center for Educational Accountability in the School of Education at The University of Alabama at Birmingham. Correspondence regarding this paper may be sent via e-mail to Dr. Wu at <ywu@TMTC760.TMTC.EDU.TW> or to Dr. McLean at <JMcLean@uab.edu>. Traditional correspondence may be sent to James E. McLean, UAB School of Education, 901 13th. Street, South, Birmingham, AL 35294-1250.

BEST COPY AVAILABLE

recommendations by Keppel (1973) and Winer (1971), may not have provided the optimal number of blocks to achieve maximum power. This potential limitation is magnified by the restricted range of experimental conditions they used.

Wu and McLean (1994b) examined the blocking versus ANCOVA issue and estimated the optimal number of blocks to achieve maximum power using broader and more representative experimental conditions. They recommended that, when deciding among a completely randomized ANOVA, a block design, or ANCOVA, researchers should consider the assumptions of the procedures and weigh the magnitude of the power increase against the added cost of blocking or covarying. The Wu and McLean study also described how Apparent Imprecision compares with statistical power and why the correlation between the concomitant and dependent variables should be considered the critical factor in choosing blocking or ANCOVA based on Apparent Imprecision (see Appendix A). They pointed out the potential problems with Apparent Imprecision and recommended power be used in preference to Apparent Imprecision (see Appendix A). Generally, the Wu and McLean study supported the recommendation by Maxwell and Delaney (1984) to use ANCOVA in preference to blocking if the assumptions for ANCOVA can be met. Nevertheless, results of the Wu and McLean study concluded that ANCOVA is not always more powerful than blocking, as suggested by Maxwell and Delaney. However, the Wu and McLean study was limited to using only post-hoc assignment of the concomitant variable.

The Maxwell and Delaney study (1984) explored another dimension to the blocking versus ANCOVA issue by using the concomitant variable to assign subjects to treatments. If the concomitant variable is not considered when subjects are assigned to treatments, the experiment uses a post-hoc approach (Bonett, 1982; Keppel, 1973; Myers, 1979); otherwise an a priori approach is used. Maxwell and Delaney (1984) found that ANCOVA tends to be more powerful than blocking if the same approach is selected, and a priori assignment tends to be more powerful than post-hoc if the same procedure is selected. The purpose of the present study is to compare the statistical powers of a priori and post-hoc approaches among ANOVA, block designs, and ANCOVA using broad, representative experimental conditions and varying the numbers of blocks based on statistical power.

Procedures

Experimental Conditions

This Monte Carlo study compares the statistical powers among ANOVA, block designs, and ANCOVA under 48 experimental conditions with both post-hoc and a priori approaches. The 48 experimental conditions are combinations of four levels of the number of treatments (T; 2, 3, 4, 5), three levels of the number of subjects per treatment (n; 8, 40, 72), and four levels of the correlation coefficient between the concomitant and dependent variables (ρ; .00, .28, .56, .84). The levels of experimental conditions were selected to achieve equal intervals and to be representative of real world situations. The four levels of the number of treatments represent the most commonly used numbers of treatments; the three levels of the number of subjects per treatment represent small, medium, and large sample sizes; and the four levels of the correlation coefficient represent zero, low, moderate, and high correlations.

Method of Assignment

For the a priori approach, the required total number of subjects was randomly selected and ranked by the magnitude of the concomitant variables. The highest ranked k subjects formed the first block; the second highest ranked k subjects formed the second block; and so on until the lowest ranked k subjects formed the nth block, where k is the number of treatments and n is the number of subjects per treatment. The subjects in each block then were randomly assigned to treatments. This method was chosen because of its simplicity and its ability to form the most homogeneous blocks. With this method of assignment, the concomitant variable was classified as a priori based on Maxwell and Delaney (1984). For the post-hoc approach, subjects were randomly assigned to treatments without considering the concomitant variable.

Methods of Analysis

For ANOVA, the concomitant variable was not considered in the analysis. Thus, the ANOVA served as the control condition. For block designs, subjects in each treatment were blocked by their ranks on the concomitant variable. The block analyses included all possible numbers of blocks to achieve equal numbers of subjects in each block. For example, with 72 subjects per treatment, analyses were conducted with 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, and 72 blocks. For ANCOVA, the concomitant variable was treated as the covariate in the analysis.

Post-hoc ANOVA (Completely Randomized ANOVA) as the Control Group

The experiment controlled the power of the post-hoc ANOVA (completely randomized ANOVA) at .50 using the effect sizes reported by Wu (1994). This allows the power of the other procedures to increase or decrease as a function of experimental conditions and to make comparisons of the analysis procedures more meaningful with the post-hoc ANOVA serving as a control group.

Computer Simulation System

This study used a Monte Carlo approach so-named because it uses data simulated according to a prescribed set of conditions (Jain, 1992) based on probabilities of certain events occurring. This approach is well-documented (e.g., Jain, 1992) and provides researchers with a method of simulating almost any desirable data condition. It would be very difficult if not impossible to find existing data that met all of the conditions that the researcher wanted to compare. Since the researcher sets the conditions before simulation, the researcher knows the values of the population parameters. This allows the researchers to compare the results of an analysis with the true population values and judge the accuracy of the analysis. This is best accomplished by generating thousands of sets of data and replicating the analysis or analyses with each set of data. Monte Carlo simulation methods have been used successfully in a number of methodological statistical studies where analytic methods were not feasible (e.g., Maxwell & Delaney, 1984; McLean, 1974; Neel, 1970; Wu & McLean, 1994b).

A Monte Carlo computer simulation system developed by Wu and McLean (1993) and used by Wu (1994) and Wu and McLean (1994b) was modified for this study. This computer simulation system has been demonstrated to be capable of generating data that meet predetermined specifications and carrying out accurate simulations; also, the computer programs can be modified easily for many other studies (Wu & McLean, 1994a). For this study, paired data were generated from two linearly correlated normal populations. Data generated in this nature will meet the assumptions of ANOVA and ANCOVA but will not completely satisfy the assumptions of block designs. Robustness of block designs under the circumstances has been illustrated analytically by Feldt (1958) and empirically by Maxwell and Delaney (1984).

The computer programs used by Wu (1994) and Wu and McLean (1994b) were used for the post-hoc approach in this study. The specific computer codes and a detailed description of the simulation procedures can be found in Wu. The same seed numbers used in the two previous studies were used in this study. The results demonstrated Wu's statement that experiments are replicable using the same seeds. The computer programs for the a priori approach are slightly different from the post-hoc approach and are listed in Appendix B. The same seeds used for the post-hoc approach were used for the a priori approach in order to compare the two approaches based on analyzing the same sets of data.

Results

The resulting power values under each experimental condition are listed in Appendix C. Each power is based on the analysis of 3,000 sets of data with alpha preset at .05.

Optimal Number of Blocks

Results show that the optimal number of blocks to achieve maximum power increases as the correlation, the number of treatments, and the number of subjects per treatment increase. The optimal number of blocks for the a priori approach is essentially the same as that for the post-hoc approach. As adjacent numbers of blocks often yield very close power values, selection of the optimal number of blocks has no clear-cut boundary. In fact, the power values are almost the same as the number of blocks approaches its optimal number. This is important for researchers because the optimal number of blocks may not be as crucial as it has been regarded, although theoretically there exists an optimal number of blocks. The results of this study support the recommendation by Wu and McLean (1994b) that researchers may select from a wide range of numbers of blocks if they avoid using small numbers of blocks when the correlation, the number of treatments, and the number of subjects per treatment are large, and vice versa. The optimal numbers of blocks are listed in Table 1. The purpose of this table is to demonstrate the trend that the optimal number of blocks increases as the correlation, the number of treatments, and the number of subjects per treatment increase, rather than provide strict decision rules for selecting the optimal numbers of blocks.

Post-hoc ANOVA (Completely Randomized ANOVA) as the Control Group

When the correlation coefficient is zero, both a priori and post-hoc ANOVAs are as powerful or more powerful than blocking and ANCOVA. They are more powerful when the number of subjects per treatment (n) and the number of treatments (T), especially n, are small. This is plausible because blocking and ANCOVA achieve no advantage over a completely randomized ANOVA when the correlation is zero and using blocking or ANCOVA reduces the degrees of freedom available to estimate error. Thus, for block designs, power is diminished when a larger number of blocks is used. The loss of degrees of freedom has little impact when the sample size is large, but causes significantly more power loss as the sample size is reduced.

Table 1
The Optimal Number of Blocks to Achieve Statistical Power

Correlation Coefficient*	Number of Treatments	Number of Subjects per Treatment 8	40	72
.28	2	2	10	18
	3	4	10	24
	4	4	20	24
	5	4	20	24
.56	2	4	10	24
	3	4	20	36
	4	8	20	36
	5	8	20	36
.84	2	4	20	36
	3	8	20	36
	4	8	40	72
	5	8	40	72

*Between concomitant and dependent variables.

The power of post-hoc ANOVA (completely randomized ANOVA) is controlled at .50 under all experimental conditions. The power of a priori ANOVA is also controlled at .50 when the correlation is zero. This is also plausible because a priori ANOVA is no different from the completely randomized ANOVA when the correlation is zero. But the power of a priori ANOVA drops as ρ and T increase. For example, a priori ANOVA loses .27 power with T=5, *n*=72, and ρ=.84. The magnitudes of the power losses under each experimental condition are listed in Table 2. Losses of less than .02 are omitted for clarity.

Power of Block Designs and ANCOVA

The completely randomized ANOVA is the best choice and there is no need to block or covary when the correlation is zero. When the correlation is not zero, blocking and ANCOVA become more powerful than the completely randomized ANOVA as ρ, *n*, and T increase. Neither one procedure nor one approach is uniformly most powerful. ANCOVA is not generally more powerful than blocking, and the a priori approach is not generally more powerful than the post-hoc approach. The relative merits of the procedures are complicated, and the choice of the optimal procedure varies depending on the experimental conditions. The power increases for both blocking and ANCOVA are listed in Table 3 and 4.

Table 2
Power Difference Between A Priori and Post-hoc ANOVA

Number of Subjects per Treatment	Correlation Coefficient*	Number of Treatments 2	3	4	5
8	.28	†	†	†	-.03
	.56	†	-.04	-.07	-.09
	.84	†	-.11	-.21	-.24
40	.28	†	†	†	-.03
	.56	†	-.03	-.06	-.10
	.84	†	-.12	-.21	-.26
72	.28	†	†	†	-.02
	.56	†	-.04	-.07	-.08
	.84	†	-.11	-.21	-.27

* Between concomitant and dependent variables.
† Denotes difference is less than .02.

Note that, under each experimental condition, all procedures analyze the same sets of data with the post-hoc ANOVA (completely randomized ANOVA) serving as the control group. The effect size of treatments is set at a specific value under each experimental condition to control the power of the completely randomized ANOVA at .50. Thus, the increase is calculated by subtracting the power of the completely randomized ANOVA from the powers of the optimal blocking procedure and ANCOVA under each experimental condition. When the correlation is low (ρ =. 28), the increases do not exceed .05 for either approach. When the correlation is moderate (ρ =. 56), the increases range from .10 to .21. When the correlation is high (ρ = .84), the increases range from .24 to .49.

Comparing Blocking and ANCOVA

The power differences between optimal blocking and ANCOVA for both approaches are listed in Tables 5 and 6.

Table 3
Power Increase Using Optimal Blocking and ANCOVA for the A Priori Approach

		Number of Treatments							
		2		3		4		5	
n*	ρ**	Optimal Block	ANCOVA	Optimal Block	ANCOVA	Optimal Block	ANCOVA	Optimal Block	ANCOVA
8	.28	†	.02	†	.02	†	.04	†	.02
	.56	.13	.17	.12	.16	.14	.17	.15	.17
	.84	.37	.46	.43	.48	.43	.46	.45	.48
40	.28	.04	.04	.03	.03	.04	.04	.03	.03
	.56	.16	.17	.18	.18	.17	.17	.17	.17
	.84	.45	.46	.46	.47	.46	.46	.48	.48
72	.28	.04	.03	.05	.05	.04	.04	.03	.03
	.56	.16	.16	.16	.16	.17	.16	.18	.18
	.84	.44	.44	.46	.46	.46	.47	.48	.48

* Denotes number of subjects per treatment.
**Denotes correlation coefficient between concomitant and dependent variables.
† Denotes difference is less than .02.

Table 4
Power Increase Using Optimal Blocking and ANCOVA for the Post-hoc Approach

		Number of Treatments							
		2		3		4		5	
n*	ρ**	Optimal Block	ANCOVA	Optimal Block	ANCOVA	Optimal Block	ANCOVA	Optimal Block	ANCOVA
8	.28	†	†	.02	†	.03	†	.04	†
	.56	.10	.13	.12	.14	.16	.17	.17	.18
	.84	.24	.43	.32	.46	.36	.45	.40	.48
40	.28	.03	.03	.03	.03	.05	.04	.05	.03
	.56	.13	.16	.16	.17	.19	.17	.21	.18
	.84	.31	.46	.38	.47	.41	.46	.45	.48
72	.28	.03	.03	.04	.04	.05	.03	.05	.03
	.56	.14	.16	.16	.15	.20	.19	.21	.17
	.84	.30	.44	.37	.46	.42	.46	.45	.49

* Denotes number of subjects per treatment.
**Denotes correlation coefficient between concomitant and dependent variables.
† Denotes difference is less than .02.

BEST COPY AVAILABLE

Table 5
Power Differences Between Optimal Blocking and ANCOVA for the A Priori Approach

		Number of Treatments			
*n**	ρ**	2	3	4	5
	.28	†	†	†	†
8	.56	.04	.04	.04	.02
	.84	.09	.04	.03	.02
	.28	†	†	†	†
40	.56	†	†	†	†
	.84	†	†	†	†
	.28	†	†	†	†
72	.56	†	†	†	†
	.84	†	†	†	†

* Denotes number of subjects per treatment.
** Denotes correlation coefficient between concomitant and dependent variables.
† Denotes difference is less than .02.

Table 6
Power Differences Between Optimal Blocking and ANCOVA for the Post-hoc Approach

		Number of Treatments			
*n**	ρ**	2	3	4	5
	.28	†	†	†	-.02
8	.56	.03	†	†	†
	.84	.19	.14	.10	.08
	.28	†	†	†	-.02
40	.56	.03	†	†	-.03
	.84	.14	.09	.05	.03
	.28	†	†	†	-.02
72	.56	†	†	†	-.04
	.84	.15	.09	.04	.04

* Denotes number of subjects per treatment.
** Denotes correlation coefficient between concomitant and dependent variables.
† Denotes difference is less than .02.

For the a priori approach, ANCOVA is more powerful than the optimal blocking procedure when the number of subjects per treatment is small and the correlation is moderate or high. For the post-hoc approach, ANCOVA is more powerful when the correlation is high, while the optimal blocking procedure is slightly more powerful when the correlation is low or moderate and the number of treatments is large. These findings are different from those based on Apparent Imprecision that suggest ANCOVA is consistently better than blocking as the correlation increases (Feldt, 1958). Rather, the results support Maxwell and Delaney's (1984) statement that "the recommendation of most experimental design texts to consider the correlation between the dependent and concomitant variables in choosing the best technique for utilizing a concomitant variable is incorrect" (p. 136).

Comparing A Priori and Post-hoc Approaches

Overall, the a priori approach yields little advantage over the post-hoc approach. The power means of the a priori and post-hoc approaches of all the analysis procedures over all the experimental conditions except those with zero correlation are listed in Table 7. Tables 8 and 9 show the power differences between a priori and post-hoc approaches for optimal blocking and ANCOVA.

Table 7
Mean Powers of Analysis Procedures for A Priori and Post-hoc Approaches

	Approach	
Design	a priori	post hoc
ANOVA	.432	.500
2 blocks	.630	.629
3 blocks	.679	.665
4 blocks	.689	.671
5 blocks	.705	.684
6 blocks	.708	.690
8 blocks	.703	.680
9 blocks	.715	.696
10 blocks	.717	.695
12 blocks	.717	.698
18 blocks	.719	.700
20 blocks	.721	.699
24 blocks	.720	.701
36 blocks	.721	.702
40 blocks	.719	.696
72 blocks	.720	.700
ANCOVA	.723	.717

A priori blocking is more powerful than post-hoc blocking when the correlation is high. Post-hoc blocking is slightly more powerful than a priori blocking when the correlation is low or moderate and the number of treatments is large. A priori ANCOVA is more powerful than post-hoc ANCOVA when the number of subjects per treatment and the number of treatments are small. These results do not support Maxwell and Delaney's (1984) conclusion that the a priori approach is generally more powerful than the post-hoc approach. However, this study does support their findings for similar experimental conditions. But, Maxwell and Delaney used a much narrower set of experimental conditions. Specifically, the power values in the upper-left three cells (T=2, n=8, and ρ=.28, 56, and 84) of Tables 8 and 9 actually support the results reported by Maxwell and Delaney. The pattern of the magnitudes of power differences is analogous to that of the Maxwell and Delaney study, where the magnitudes of differences are generally small except the one between a priori and post-hoc blocking when the correlation is high and the number of subjects per treatment and the number of treatments are small.

Table 8
Power Differences Between A Priori and Post-hoc Approaches for Optimal Blocking

Number of Subjects per Treatment	Correlation Coefficient*	Number of Treatments			
		2	3	4	5
8	.28	†	†	†	-.03
	.56	.03	†	-.03	-.02
	.84	.13	.12	.07	.05
40	.28	†	†	†	-.02
	.56	.04	†	-.02	-.04
	.84	.14	.08	.05	.03
72	.28	†	†	†	-.02
	.56	.02	†	-.04	-.03
	.84	.14	.09	.05	.04

* Between concomitant and dependent variables.
† Denotes difference is less than .02.

Table 9
Power Differences Between A Priori and Post-hoc Approaches for ANCOVA

Number of Subjects per Treatment	Correlation Coefficient*	Number of Treatments			
		2	3	4	5
8	.28	.03	†	†	†
	.56	.04	.02	†	†
	.84	.02	†	†	†
40	.28	†	†	†	†
	.56	†	†	†	†
	.84	†	†	†	†
72	.28	†	†	†	†
	.56	†	†	-.03	†
	.84	†	†	†	†

* Between concomitant and dependent variables.
† Denotes difference is less than .02.

Discussion and Recommendations

No one procedure or single approach is uniquely more powerful. Although the most powerful technique to employ a concomitant variable varies depending on the experimental conditions, most of the magnitudes of the power differences are not large enough to be practically significant. It is recommended that researchers utilize the tables provided in this study to help select the best technique when employing a concomitant variable.

The problems concerning utilizing a concomitant variable become complicated when considering a variety of experimental conditions, methods of assignment, and assumptions of the analysis procedures. Despite these complications, the results of this study show that in research practice choices may have little impact because many of the power differences are small. Based on practical significance, this study suggests the simplest rule to follow is **use regular ANCOVA (post-hoc ANCOVA) if its assumptions can be met**. Rationale for recommending blocking over ANCOVA by earlier experimental design texts, such as ease of calculation and the ability to test simple effects, seems to have faded. With modern computer statistical packages, ANCOVA has become at least as easy to compute as block designs. Using regular ANCOVA, researchers need not consider questions such as whether the concomitant variable is available before the experiment, how to assign subjects, what is the magnitude of the correlation, and how many blocks should be used. They will still increase their power though not necessarily achieve maximum power. The power differences between post-hoc ANCOVA and the optimal procedure are of little practical significance under most experimental conditions and do not exceed .04 even in the most extreme cases.

When the correlation is zero, the waste of degrees of freedom due to blocking and ANCOVA may result in the reduction of power. Wasting degrees of freedom has little influence on power when T and n are large but has an effect when T and n are small. For example, the power drops from .50 to .43 when using a post-hoc eight block analysis procedure for ρ=.00, T=2 and n=8. Because this study uses a minimum n of 8, the loss of one degree of freedom using ANCOVA, especially a priori ANCOVA, when the correlation is zero, seems to have little effect on the loss of power. However, one should be cautious that the loss will increase as n becomes smaller. Much of the criticism of the post-hoc approach is based on the ease with which researchers can block or covary in a post-hoc manner. Myers (1979) pointed out the danger of abusing post-hoc block designs by demonstrating that reordering scores within each treatment does not change the treatment means but generally reduces the error variance, resulting in significant *F*s which "merely reflect the reduction in error variance due to blocking rather than any variability due to treatments" (p. 155). However, Myers did not consider the loss of degrees of freedom with block designs. Wasting degrees of freedom on some nonsense concomitant variable would simply decrease power. Nevertheless, the caution urged by Myers should be considered. It is often too easy to peek at the data, play with several concomitant variables, or try several analysis procedures to achieve significant results. However, these should be considered ethical problems rather than problems of the post-hoc approach per se. Researchers should neither block nor covary unless they can justify the concomitant variable before the analysis. If researchers would always consider practical as well as statistical significance, these problems could be avoided as none of these analysis techniques affect effect sizes.

One of the most interesting findings of this study is the problem with a priori ANOVA. Maxwell and Delaney (1984) questioned this method and detected minor Type I error rate problems with it. But the Type II error rate problem with a priori ANOVA was not detected in their study. This is because their study was limited to two treatments. Future research could investigate the Type I error rate using a broader range of experimental conditions. A follow-up Monte Carlo study comparing a priori and post-hoc ANOVA by examining the sample distributions of the variances showed that the power loss of a prior ANOVA was due to a decrease of treatment variance and an increase of error variance while stratified instead of randomized assignment was used. The Type II error rate problem with a priori ANOVA may provide the best example of how power can be different from Apparent Imprecision; the precision of a priori ANOVA is higher than for the completely randomized ANOVA but results in lower power. This may also explain why the a priori approach is not generally more powerful than the post-hoc approach. A priori does achieve more homogeneous blocks and ensures more equal covariates across treatments. But, the advantages are reduced by the loss of power due to stratified rather than random assignment. Stratified assignment is still a common practice. It is believed to guarantee fairness of treatments and avoid preexisting differences. Some suggest that stratified assignment reduces error and increases power. Based on the results of this study, if stratified assignment is used, the concomitant variable should not be ignored in the analysis. Since stratified assignment loses power due to non-random sampling but gains power because of the increase in the number of homogeneous blocks in the analysis, future researchers could investigate whether there is an optimal combination of the number of blocks used in assignment and the number of blocks used in analysis.

Wu and McLean (1994b) suggested four reasons that Maxwell and Delaney's (1984) results differed from theirs: (1) limited experimental conditions, (2) not including the optimal number of blocks, (3) including the interaction in the effects model, and (4) inaccuracy of the computer simulation. The results of this study suggest that not including the optimal number of blocks causes only minor power loss, and the inaccuracy of this computer simulation and that of the Maxwell and Delaney study is unlikely because results of the two studies support each other for similar experimental conditions. Among the four factors, the restriction of experimental conditions and including the interaction in the effects model could contribute the most to the different findings. Conclusions based on restricted conditions may limit generalization. As to the interaction factor, the two studies complement each other's findings as some researchers suggest pooling the interaction variance with the error variance when the interaction is non-significant while other researchers do not. Note that Maxwell and Delaney did not specify that the interaction was included in the effects model. Our conclusion is based on the statement "perform a two-way analysis of variance (ANOVA) utilizing levels of the concomitant variable as a factor in the design" (p. 138).

A randomized complete block design is defined as a block design in which each block within each treatment has only one observation. Lindquist (1953) used the term, treatments-by-levels design, which consists of more than one subject in a cell, to differentiate it from the randomized complete block design. The treatments-by-levels design is also called the treatments-by-blocks design (Kennedy & Bush, 1985). While the randomized complete block design usually uses an additive model because there is only one observation per cell, the treatments-by-blocks design can either use an additive or nonadditive model by excluding or including the interaction term in the effects model. The additive model is used in this study because the interaction does not exist in the population. Which model to be used should be based on researchers' subject matter knowledge and should be justified before the experiment. For example, suppose the concomitant variable is an IQ score, the dependent variable is a Scholastic Achievement Test (SAT) score, and the treatments are the teaching methods. If researchers can justify that the correlation between the IQ score and the SAT score should not be influenced by the teaching methods, the additive model should be used.

When an additive model is used, the concomitant variable is treated as a nuisance variable and the variance accounted for by the concomitant variable is nuisance variance, which is beyond the researcher's interest and is to be extracted only to reduce error and increase power. But, if the dependent variable is a computer attitude measure and the researcher cannot justify that high IQ students usually have better attitudes toward computers, disregarding the teaching methods used, the nonadditive model should be used. Under these circumstances, the concomitant variable is no longer a nuisance variable; rather, it is a factor of interest because the researcher would want to test if a teaching method is better for low IQ than for high IQ students. In this case, the block design in form, in analysis, and in interpretation is undistinguishable from a factorial design. The difference is that in a factorial design subjects are assigned to each combination (cell) of factors, while in block designs subjects in each block level are assigned to treatments and the block factor is usually intrinsic in the subjects themselves.

The question as to whether nonadditive block designs should be categorized as block or factorial designs adds some difficulty to the nomenclature of experimental designs. The alternative uses of terms such as blocking, factorial, and stratification by researchers certainly add confusion. Which model to use and whether to adhere to the original model or revise it during the analysis process should be justified in advance. In many instances, the researcher includes the interaction for convenience. If the interaction is non-significant, some researchers pool the interaction variance with the error variance in order to increase power. The issue of pooling and non-pooling is still disputed. If the interaction does not exist in the population, pooling the interaction variance with the error variance should provide a better estimate of the error term and increase power due to an increased number of degrees of freedom for the error term.

Using a nonadditive model suggests another question: Should the block levels formed by ranking be treated as random or fixed? If random, the interaction should be used as the error term to test the treatment effect (see Kirk, 1982, pp. 240-241); if fixed, the within-cell variance should be used as the error term (see Feldt, 1958; Maxwell & Delaney, 1984). Levels based on the rank of sample subjects seem to be more fixed, though not completely fixed, than random because ranking is deterministic instead of random. Calculating expected mean squares seems to be the most appropriate way to obtain the error terms, which is beyond the scope of this discussion. However, future Monte Carlo studies could block the population to obtain completely fixed block levels. Boundaries for block levels can be set based on two principles: equal proportion or equal interval (Feldt, 1958). In research practice, blocking a population is not feasible in most cases. An alternative method is to randomly select subjects, then fit them into corresponding levels. This would most likely result in unequal numbers of subjects in the blocks. To obtain an equal number of subjects in each block, researchers could continue randomly selecting subjects until the required number of subjects fit in each block level and discard those exceeding the required number of subjects.

It should be noted that the selection of any particular design should be predicated on meeting the assumptions required for that particular design. Generally, ANCOVA requires the most stringent set of assumptions when compared to oneway ANOVA or block designs. However, if these assumptions are satisfied, the increase in power often more than justifies the added complexity of the analysis.

A final word of caution is in order. No matter how powerful an analysis is, it does not change the effect size of the results. The more powerful the design that is used, the more likely one is to find statistical significance. Effect size should still be computed to consider the educational significance of the results.

References

Bonett, D. G. (1982). On post-hoc blocking. *Educational and Psychological Measurement, 42*, 35-39.

Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. *American Psychologist, 29*(1), 25-26.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika, 23*(4), 335-353.

Glass, G. & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn and Bacon.

Jain, S. (1992). *Monte Carlo simulations of disordered systems*. Singapore: World Scientific.

Kennedy, J. J., & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America.

Keppel, G. (1973). *Design and analysis: A researcher's handbook* (1st ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavior sciences* (2nd ed.). Monterey, CA: Brooks/Cole.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

Maxwell, S. E., & Delaney, H. D. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin, 95*(1), 136-147.

McLean, J. E. (1974). An empirical examination of analysis of covariance with and without Porter's adjustment for a fallible covariate. *Dissertation Abstracts International*, 36. (University Microfilms No. 579-1131A)

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn and Bacon.

Neel, J. (1970). *A comparative analysis of some measures of change.* Unpublished doctoral dissertation, University of Florida, Gainesville.

Thompson, B. (1985). Alternate methods for analyzing data from educational experiments. *Journal of Experimental Education, 54*, 50-55.

Thompson, B. (1994). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: the neo-classical perspective. In B. Thompson (Ed.), *Advances in social science methodology* (pp. 3-28, Vol. 3). Greenwich, CT: JAI Press.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Wu, Y. (1994). *To block or covary a concomitant variable: Which is more powerful?* Unpublished doctoral dissertation, The University of Alabama, Tuscaloosa, AL.

Wu, Y., & McLean, J. E. (1993, November). *To block or covary a concomitant variable: Which is better?* Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Wu, Y., & McLean, J. E. (1994a). Using SAS® for Monte Carlo Simulation. *SUGI 19 proceedings.* Cary, NC: SAS Institute Inc.

Wu, Y., & McLean, J. E. (1994b). *Choosing the optimal number of blocks based on power.* Manuscript submitted for publication.

Appendix A

Examining Apparent Imprecision may provide insight into how it is similar to or different from power. Apparent Imprecision is defined as the product of True Imprecision and an adjustment factor based on the degrees of freedom for error:

$$I_a = \frac{ave\ var}{min\ var} \times \frac{df_e+3}{df_e+1},$$

where, ave var is the average variance of the treatment mean difference from sample to sample; and min var is the theoretical minimum variance of the treatment mean difference. According to Feldt (1958), the minimum variance is the variance of the dependent variable at a fixed value of the covariate given the assumption of homoscedasticity. For block designs, Apparent Imprecision (AI_{BD}) is computed with the following formula:

$$AI_{BD} = \frac{\frac{2\sigma_y^2}{n}[1-\rho^2(1-\frac{\bar{\sigma}_x^{\,2}}{\sigma_x})]}{\frac{2\sigma_y^2}{n}(1-\rho^2)} \times \frac{f_{e_y}+3}{f_{e_y}+1} = \frac{[1-\rho^2(1-\frac{\bar{\sigma}_x^{\,2}}{\sigma_x})]}{(1-\rho^2)} \times \frac{f_{e_y}+3}{f_{e_y}+1},$$

where Y represents the dependent variable; X, the concomitant variable; ρ, the correlation coefficient between X and Y; n, the number of subjects per treatment; $\bar{\sigma}_x^{\,2}$, the average variance of X over all blocks; and f_{e_y}, the degrees of freedom for error in Y. For ANCOVA, Apparent Imprecision is computed using the following formula:

$$AI_{ANCOVA} = \frac{\frac{2\sigma_y^2}{n}(1-\rho^2)(1+\frac{1}{f_{e_x}-2})}{\frac{2\sigma_y^2}{n}(1-\rho^2)} \times \frac{f_{e_y}+3}{f_{e_y}+1} = \frac{f_{e_x}-1}{f_{e_x}-2} \times \frac{f_{e_y}+3}{f_{e_y}+1},$$

where f_{e_x} stands for the degrees of freedom for error in X.

For block designs, the average variance of the covariate over all blocks decreases and the True Imprecision approaches 1.00 as the number of blocks increases. Theoretically, if an infinite number of blocks could be used, the values of the concomitant variable would be the same in each block, and True Imprecision would be exactly 1.00. This does not mean that employing a larger number of blocks always decreases Apparent Imprecision, because as larger numbers of blocks are used, larger degrees of freedom for error are lost, and, based on the adjustment factor, Apparent Imprecision increases. Therefore, there is an optimal number of blocks that minimizes True Imprecision, yet, on the other hand, minimizes the increase of Apparent Imprecision due to the loss of degrees of freedom for error. This phenomenon of Apparent Imprecision is analogous to that of power. For the same reason, there is an optimal number of blocks such that the higher the degree of the correlation between the concomitant and dependent variables, the more homogeneous the values of the dependent variable are in each block, the more variance that is extracted from the error term, the higher the power. At the same time, the power is decreased due to the loss of degrees of freedom for error. To optimize power, one must find a balance between these two forces. Apparent Imprecision is similar to power in this regard.

Nevertheless, Apparent Imprecision could suggest different results from power. For block designs, the average variance and the minimum variance decrease as the correlation increases. But, the average variance would never decrease as much as the minimum variance unless an infinite number of blocks was used. Therefore, the Apparent Imprecision of a block design usually increases as the correlation increases. For ANCOVA, the correlation terms in the numerator and denominator are canceled out because the average variance is a function of the minimum variance. Notice that the minimum variance is the ideal variance based on the covariance model. Therefore, the Apparent Imprecision of ANCOVA is the same for all values of the correlation. This is basically why block designs are found to consistently become less precise than ANCOVA as the correlation increases and why the correlation has been regarded as the critical factor in choosing block designs or ANCOVA. The correlation being negative in a block design and irrelevant in ANCOVA based on Apparent Imprecision is different from what most texts and this study have found about the positive effect of higher correlation in reducing error and increasing power.

A simple rule to follow when evaluating a criterion variable is to determine if it provides a direct measure of the variable of interest. For example, if a new brand of bulbs is to be evaluated, it is better to check how long the bulbs last rather than to analyze the precision of the components of the bulbs. The precision of the components of the bulbs would be a good criterion if it could determine a useful property; for example, the more precise the element is, the longer the bulbs last, the less power the bulbs consume, or the less eye strain the bulbs cause. A theoretical framework merits less if the degree it can be related to the physical property of interest is low. For example, the theory of the imagined number $\sqrt{-1}$ would not have been valuable if it could not be used to predict the behavior of electronic circuits. Based on the rule, the Type I error rate and statistical power should be considered a good criterion in evaluating a research design.

Maxwell and Delaney (1984) also illustrated how Apparent Imprecision might be different from power:

> Suppose an experimenter plans to conduct a two-group comparison of means using an alpha level of 0.05. If the population difference in means is .5 standard deviation units, and 150 subjects are randomly and independently assigned to groups, the power for an independent groups *t* test is 0.99. Suppose that a concomitant variable were available that correlated .6 with the dependent variable. The power of an ANCOVA would still be 0.99 (or, actually, 1.00 if rounded to two decimal places). From the standpoint of power, the ANCOVA offers no gain over the *t* test. On the other hand, it can be shown that the apparent imprecision of the *t* test here is 1.573, whereas for ANCOVA the apparent imprecision is 1.010, demonstrating that the estimated magnitude of the treatment effect is much more precise when ANCOVA is used rather than the *t* test. (p. 137)

One might interpret, facially, the above demonstration as a benefit of using Apparent Imprecision as the criterion variable. However, it should be noticed that the powers of the *t* test and the ANCOVA have both reached the ceiling point because of the large sample size that was used. Based on the rule, this illustration, indeed, offers an example of power as a favorable criterion. If one can achieve a .99 power with a *t* test, what is the advantage of spending money on collecting concomitant data? For example, administering a pretest or IQ test, to gain an impractical .001 power. Eventually, almost all analysis will become statistically significant if a large enough size is used. What is the use of a new teaching method that claims to increase students' SAT scores by 1 point? Practical significance would also need to be considered when evaluating the results of an analysis.

Appendix B

Executable File

```
/*   */
ADDRESS COMMAND
"ERASE PVALUE DATA A"
NUMERIC DIGITS 10
TIME = 1
DO WHILE TIME < 1001
SEED = 2132560 + (TIME -1) * 2147483
"EXECIO 1 DISKW" NEWSEED DATA A "(STRING" SEED
"EXEC SAS T57284"
"ERASE NEWSEED DATA A"
TIME=TIME+1
END
"EXEC SAS T57284P"
```

First SAS Program (T57284 SAS A)

```
CMS FILEDEF INDATA DISK NEWSEED DATA A;
CMS FILEDEF PVALUE DISK PVALUE DATA A (LRECL
306 BLKSIZE 306 RECFM FBS;
CMS FILEDEF SASLIST DISK T57284 LISTING A;
DATA BIVNORM;
  INFILE INDATA;
    INPUT SEED;
  DO I=1 TO 360;
    X=RANNOR(SEED);
    Y=.84*X+SQRT(1-.84**2)*RANNOR(SEED);
    OUTPUT;
  END;
PROC SORT;
  BY X;
DATA BIVNORM;
  SET BIVNORM;
  B72=CEIL(_N_/5);
  B2=CEIL(B72/36);B3=CEIL(B72/24);
  B4=CEIL(B72/18);B6=CEIL(B72/12);
  B8=CEIL(B72/9);B9=CEIL(B72/8);
  B12=CEIL(B72/6);B18=CEIL(B72/4);
  B24=CEIL(B72/3);B36=CEIL(B72/2);
PROC SORT;
  BY B72 I;
DATA BIVNORM (DROP=SEED I);
  SET BIVNORM;
  GROUP=MOD(_N_,5); IF GROUP=0 THEN GROUP=5;
  IF GROUP=2 THEN Y=0.0951+Y;
  IF GROUP=3 THEN Y=0.1903+Y;
  IF GROUP=4 THEN Y=0.2854+Y;
  IF GROUP=5 THEN Y=0.3805+Y;
PROC PRINT;
PROC SORT;
  BY GROUP;
PROC CORR DATA=BIVNORM;
  VAR X Y;
  BY GROUP;
PROC ANOVA;
  CLASS GROUP;
  MODEL Y=GROUP;


PROC ANOVA;
  CLASS GROUP B2;
  MODEL Y=GROUP B2;
PROC ANOVA;
  CLASS GROUP B3;
  MODEL Y=GROUP B3;
PROC ANOVA;
  CLASS GROUP B4;
  MODEL Y=GROUP B4;
PROC ANOVA;
  CLASS GROUP B6;
  MODEL Y=GROUP B6;
PROC ANOVA;
  CLASS GROUP B8;
  MODEL Y=GROUP B8;
PROC ANOVA;
  CLASS GROUP B9;
  MODEL Y=GROUP B9;
PROC ANOVA;
  CLASS GROUP B12;
  MODEL Y=GROUP B12;
PROC ANOVA;
  CLASS GROUP B18;
  MODEL Y=GROUP B18;
PROC ANOVA;
  CLASS GROUP B24;
  MODEL Y=GROUP B24;
PROC ANOVA;
  CLASS GROUP B36;
  MODEL Y=GROUP B36;
PROC ANOVA;
  CLASS GROUP B72;
  MODEL Y=GROUP B72;
PROC GLM;
  CLASS GROUP;
  MODEL Y=GROUP X/SS3;
DATA;
  INFILE SASLIST;
  INPUT WORD1 $ WORD2 $ @;
  FILE PVALUE MOD;
  IF WORD1 = 'X' AND WORD2 ='72' THEN DO;
    INPUT MEAN STDDEV;
    PUT MEAN 6.4 STDDEV 6.4 @;
    INPUT Y $ N MEAN STDDEV;
    PUT MEAN 6.4 STDDEV 6.4 @;
    END;
  ELSE  IF WORD1="X" AND WORD2 = '1.00000' THEN
  DO;
    INPUT CORR;
    PUT CORR 6.4 @;
    END;
```

BEST COPY AVAILABLE

```
  ELSE  IF WORD1="GROUP" AND WORD2 = '4' THEN
DO;
    INPUT SS MS F PR;
    PUT PR 6.4 @;
    INPUT BLOCK $ DF SS MS F PR;
    PUT PR 6.4 @;
    END; */
```

<u>Second SAS Program (T57284P SAS A)</u>

```
CMS FILEDEF INDATA DISK PVALUE DATA A;
DATA PVALUE;
INFILE INDATA;
INPUT (G1XMEAN G1XSD G1YMEAN G1YSD G1CORR
      G2XMEAN G2XSD G2YMEAN G2YSD G2CORR
      G3XMEAN G3XSD G3YMEAN G3YSD G3CORR
      G4XMEAN G4XSD G4YMEAN G4YSD G4CORR
      G5XMEAN G5XSD G5YMEAN G5YSD G5CORR
      GROUP1B BLOCK1B GROUP2B BLOCK2B GROUP3B
      BLOCK3B GROUP4B BLOCK4B GROUP6B BLOCK6B
            GROUP8B  BLOCK8B  GROUP9B  BLOCK9B
GROUP12B
          BLOCK12B  GROUP18  GROUP18B  BLOCK18B
GROUP24B
      BLOCK24B GROUP36B BLOCK36B GROUP72B
      BLOCK72B GROUPANC BLOCKANC) (51* 6.4);
    G1BSG=0;
    B1BSG=0;
    G2BSG=0;
    B2BSG=0;
    G3BSG=0;
    B3BSG=0;
    G4BSG=0;
    B4BSG=0;
    G6BSG=0;
    B6BSG=0;
    G8BSG=0;
    B8BSG=0;
    G9BSG=0;
    B9BSG=0;
    G12BSG=0;
    B12BSG=0;
    G18BSG=0;
    B18BSG=0;
    G24BSG=0;
    B24BSG=0;
    G36BSG=0;
    B36BSG=0;
    G72BSG=0;
    B72BSG=0;
    GANCSG=0;
    BANCSG=0;
    TOTAL=1;
    IF GROUP1B <= 0.05 THEN G1BSG=1;
    IF BLOCK1B <= 0.05 THEN B1BSG=1;
    IF GROUP2B <= 0.05 THEN G2BSG=1;
    IF BLOCK2B <= 0.05 THEN B2BSG=1;
    IF GROUP3B <= 0.05 THEN G3BSG=1;
    IF BLOCK3B <= 0.05 THEN B3BSG=1;
    IF GROUP4B <= 0.05 THEN G4BSG=1;
    IF BLOCK4B <= 0.05 THEN B4BSG=1;
    IF GROUP6B <= 0.05 THEN G6BSG=1;
    IF BLOCK6B <= 0.05 THEN B6BSG=1;
    IF GROUP8B <= 0.05 THEN G8BSG=1;
    IF BLOCK8B <= 0.05 THEN B8BSG=1;
    IF GROUP9B <= 0.05 THEN G9BSG=1;
    IF BLOCK9B <= 0.05 THEN B9BSG=1;
    IF GROUP12B <= 0.05 THEN G12BSG=1;
    IF BLOCK12B <= 0.05 THEN B12BSG=1;
    IF GROUP18B <= 0.05 THEN G18BSG=1;
    IF BLOCK18B <= 0.05 THEN B18BSG=1;
    IF GROUP24B <= 0.05 THEN G24BSG=1;
    IF BLOCK24B <= 0.05 THEN B24BSG=1;
    IF GROUP36B <= 0.05 THEN G36BSG=1;
    IF BLOCK36B <= 0.05 THEN B36BSG=1;
    IF GROUP72B <= 0.05 THEN G72BSG=1;
    IF BLOCK72B <= 0.05 THEN B72BSG=1;
    IF GROUPANC <= 0.05 THEN GANCSG=1;
    IF BLOCKANC <= 0.05 THEN BANCSG=1;
PROC FREQ;
    TABLE G1BSG -- BANCSG;
PROC SUMMARY DATA=PVALUE;
    VAR G1XMEAN -- BANCSG;
    OUTPUT OUT = DESCRIPT;
PROC PRINT DATA=DESCRIPT;
PROC UNIVARIATE DATA=PVALUE PLOT NORMAL;
    VAR G1XMEAN -- BLOCKANC;
```

Appendix C

Power Table

				ANO	B02	B03	B04	B05	B06	B08	B09	B10	B12	B18	B20	B24	B36	B40	B72	COV
T2	n08	C00	A	.505	.498		.489			.449										.498
			P	.497	.491		.481			.430										.457
		C28	A	.506	.519		.513			.472										.527
			P	.505	.510		.511			.470										.501
		C56	A	.500	.592		.616			.584										.658
			P	.490	.565		.589			.554										.623
		C84	A	.484	.773		.872			.870										.958
			P	.501	.680		.740			.730										.934
	n40	C00	A	.502	.498		.501	.501		.500		.500			.498			.483		.501
			P	.503	.503		.503			.502										.506
		C28	A	.508	.526		.535	.537		.533		.536			.535			.525		.539
			P	.501	.521		.529	.529		.532		.532			.527			.521		.533
		C56	A	.499	.603		.648	.655		.660		.663			.661			.652		.671
			P	.499	.583		.613	.614		.622		.623			.624			.621		.657
		C84	A	.498	.820		.908	.919		.939		.943			.949			.948		.953
			P	.498	.693		.769	.779		.797		.802			.812			.805		.954
	n72	C00	A	.502	.503	.502	.503		.502	.503	.500		.501	.500		.500	.500		.493	.503
			P	.513	.511		.511			.509										.507
		C28	A	.505	.521	.526	.530		.532	.533	.535		.535	.538		.537	.536		.539	.531
			P	.501	.522	.527	.529		.533	.531	.532		.533	.532		.531	.532		.529	.535
		C56	A	.495	.598	.622	.634		.643	.649	.649		.653	.656		.660	.660		.657	.660
			P	.499	.579	.606	.614		.624	.627	.628		.630	.629		.636	.633		.628	.655
		C84	A	.496	.819	.884	.910		.930	.939	.940		.943	.948		.947	.948		.947	.950
			P	.510	.694	.748	.769		.788	.794	.799		.805	.807		.808	.809		.806	.954
T3	n08	C00	A	.498	.495		.491			.471										.492
			P	.508	.508		.495			.481										.485
		C28	A	.483	.505		.506			.495										.516
			P	.493	.514		.516			.501										.512
		C56	A	.463	.584		.621			.619										.658
			P	.498	.591		.617			.607										.634
		C84	A	.376	.762		.888			.924										.966
			P	.490	.720		.797			.809										.952

(continued)

BEST COPY AVAILABLE

				ANO	B02	B03	B04	B05	B06	B08	B09	B10	B12	B18	B20	B24	B36	B40	B72	COV
	n40	C00	A	.499	.499		.498	.500		.497		.498			.498			.498		.500
			P	.503	.501		.501			.500										.498
		C28	A	.493	.525		.531	.535		.538		.536			.534			.534		.539
			P	.508	.526		.533	.535		.537		.537			.539			.537		.540
		C56	A	.469	.598		.650	.654		.671		.669			.676			.670		.678
			P	.500	.602		.641	.647		.656		.658			.662			.658		.667
		C84	A	.378	.800		.913	.927		.943		.949			.954			.956		.965
			P	.493	.744		.831	.843		.864		.864			.872			.870		.960
	n72	C00	A	.486	.486	.487	.487		.488	.485	.486		.487	.488		.484	.482		.485	.486
			P	.504	.505		.506			.504										.506
		C28	A	.494	.524	.533	.533		.536	.538	.538		.538	.538		.538	.536		.537	.544
			P	.490	.516	.523	.525		.527	.529	.528		.531	.533		.534	.533		.531	.530
		C56	A	.466	.589	.622	.636		.650	.654	.656		.658	.659		.663	.663		.661	.666
			P	.508	.610	.638	.648		.661	.665	.666		.669	.672		.670	.671		.671	.658
		C84	A	.395	.815	.902	.929		.945	.952	.953		.957	.961		.961	.964		.964	.964
			P	.504	.742	.807	.834		.851	.858	.863		.867	.871		.875	.875		.873	.968
T4	n08	C00	A	.504	.505		.497			.484										.502
			P	.497	.496		.497			.477										.485
		C28	A	.488	.516		.519			.510										.536
			P	.501	.529		.532			.522										.518
		C56	A	.427	.572		.619			.632										.669
			P	.495	.616		.651			.657										.666
		C84	A	.305	.752		.897			.935										.966
			P	.510	.764		.853			.865										.964
	n40	C00	A	.509	.508		.508	.506		.507		.507			.504			.503		.510
			P	.505	.507		.506			.504										.500
		C28	A	.479	.516		.524	.525		.528		.526			.529			.529		.532
			P	.494	.523		.536	.536		.540		.536			.541			.535		.532
		C56	A	.438	.581		.639	.648		.656		.659			.668			.668		.674
			P	.502	.625		.662	.669		.679		.681			.688			.685		.672
		C84	A	.291	.787		.922	.940		.953		.958			.962			.964		.969
			P	.505	.782		.875	.888		.905		.908			.913			.914		.968

(continued)

BEST COPY AVAILABLE

				ANO	B02	B03	B04	B05	B06	B08	B09	B10	B12	B18	B20	B24	B36	B40	B72	COV
	n72	C00	A	.503	.503	.501	.502		.499	.502	.501		.500	.502		.503	.501		.504	.502
			P	.506	.508		.508			.508										.508
		C28	A	.483	.510	.517	.520		.521	.524	.525		.524	.528		.525	.526		.524	.528
			P	.493	.529	.533	.537		.539	.539	.541		.540	.541		.541	.543		.543	.527
		C56	A	.428	.578	.611	.630		.641	.651	.655		.657	.661		.661	.663		.661	.660
			P	.498	.626	.662	.677		.686	.693	.695		.695	.698		.699	.701		.699	.692
		C84	A	.303	.811	.900	.930		.954	.959	.961		.965	.965		.967	.969		.970	.973
			P	.508	.788	.849	.877		.901	.912	.915		.920	.923		.923	.925		.924	.969
T5	n08	C00	A	.515	.512		.511			.496										.509
			P	.508	.503		.502			.488										.486
		C28	A	.471	.507		.515			.514										.522
			P	.502	.532		.540			.533										.518
		C56	A	.421	.584		.646			.658										.679
			P	.510	.633		.678			.680										.689
		C84	A	.252	.757		.906			.947										.971
			P	.494	.791		.875			.893										.970
	n40	C00	A	.504	.504		.505	.503		.500		.501			.502			.502		.504
			P	.509	.509		.511			.513										.506
		C28	A	.479	.515		.526	.530		.530		.532			.529			.530		.537
			P	.504	.537		.549	.551		.555		.555			.554			.552		.532
		C56	A	.413	.582		.641	.652		.666		.668			.678			.678		.680
			P	.508	.645		.690	.702		.710		.711			.714			.712		.685
		C84	A	.238	.781		.926	.942		.960		.966			.973			.973		.979
			P	.495	.814		.901	.913		.931		.934			.940			.940		.973
	n72	C00	A	.502	.503	.502	.501		.502	.500	.501		.500	.503		.501	.502		.501	.502
			P	.498	.498		.497			.495										.500
		C28	A	.474	.510	.517	.522		.524	.526	.528		.528	.529		.529	.528		.526	.531
			P	.498	.532	.541	.544		.547	.547	.549		.548	.548		.549	.547		.545	.528
		C56	A	.420	.583	.627	.645		.666	.671	.673		.676	.679		.681	.683		.681	.685
			P	.504	.642	.674	.690		.703	.706	.711		.711	.714		.714	.715		.717	.675
		C84	A	.227	.781	.888	.920		.949	.961	.964		.968	.972		.972	.973		.974	.976
			P	.492	.807	.870	.895		.918	.923	.927		.930	.934		.936	.937		.936	.981

Note. Tx represents a number of treatments of x. Nxx represents a number of subjects per treatment of xx. Cxx represents a correlation coefficient between the concomitant and dependent variables of xx. ANO represents ANOVA. Bxx represents a block design with xx blocks. COV represents ANCOVA. A represents the a priori approach. P represents the post-hoc approach.

BEST COPY AVAILABLE

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◆

Please enter a subscription to Research in the Schools for:

Name: ____________________

Institution: ____________________

Address: ____________________

____

			COST
Individual Subscription ($25 per year)	Number of years	_____	_____
Institutional Subscription ($30 per year)	Number of years	_____	_____
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	_____	_____
TOTAL COST:			_____

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
RESEARCH IN THE SCHOOLS
The University of Alabama at Birmingham
School of Education, 233 Educ. Bldg.
901 13th Street, South
Birmingham, AL 35294-1250

Please note that a limited number of copies of Volumes 1 and 2 are available and can be purchased for the same subscription prices noted above.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

MSERA Membership/Renewal Form

(Please print or type)

NAME: ______________________________

TITLE: ______________________________

INSTITUTION: ______________________________

MAILING ADDRESS: ______________________________
____
____

PHONE: ____________________ FAX ____________________

ELECTRONIC MAIL ADDRESS: ______________________________

MSERA MEMBERSHIP: New___ Renewal ___
ARE YOU A MEMBER OF AERA? Yes___ No___
WOULD YOU LIKE INFORMATION ON AERA MEMBERSHIP? Yes___ No___

DUES:	Professional	$15.00	______
	Student	$10.00	______

VOLUNTARY TAX DEDUCTIBLE CONTRIBUTION TO MSER FOUNDATION ______

TOTAL ______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. John Enger
Arkansas State University
P.O. Box 2535
State University, AR 72401

Research in the Schools
Mid-South Educational Research Association
and The University of Alabama
Post Office Box 870231
Tuscaloosa, AL 35487-0231

NON-PROFIT ORG.
U.S. POSTAGE
PAID
TUSCALOOSA, AL
PERMIT NO. 16

TM-CISE

MSERA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and the University of Alabama at Birmingham.

Volume 3, Number 2 **Fall 1996**

Are Students Overly Confident in Their Mathematical Errors? 1
Talia Ben-Zeev

Effects of Learning Style Accommodation on Achievement of Second Graders 9
Carol Bugg Knight, Gerald Halpin, and Glennelle Halpin

Differences in Reading and Math Achievement Test Scores for Students Experiencing Academic Difficulty 15
John R. Slate and Craig H. Jones

Preservice Teachers in Two Different Multicultural Field Programs: The Complex Influences of School Context 23
Janet C. Richards, Ramona C. Moore, and Joan P. Gipe

Beam Me Up Scottie: Professors' and Students' Experiences With Distance Learning 35
Neelam Kher-Durlabhji and Lorna J. Lacina-Gifford

The Demise of The Georgia Teacher Performance Assessment Instrument 41
Dixie McGinty

Responses That May Indicate Nonattending Behaviors in Three Self-Administered Educational Surveys. 49
J. Jackson Barnette

Influences on and Limitations of Classical Test Theory Reliability Estimates 61
Margery E. Arnold

James E. McLean and Alan S. Kaufman, Editors

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1996 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 3, Number 2 **Fall 1996**

EDITORS
James E. McLean, *University of Alabama at Birmingham*
and Alan S. Kaufman, *Psychological Assessment Resources, Inc. (PAR)*

PRODUCTION EDITOR
Margaret L. Rice, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Gypsy A. Abbott, *University of Alabama at Birmingham*
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Rhode Island*
Donald F. DeMoulin, *University of Memphis*
R. Tony Eichelberger, *University of Pittsburgh*
Daniel Fasko, Jr., *Morehead State University*
Ann T. Georgian, *Hattiesburg (Mississippi) High School*
Tracy Goodson-Espy, *University of North Alabama*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
JinGyu Kim, *National Board of Educational Evaluation (Korea)*
Jwa K. Kim, *Middle Tennessee State University*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Jerry G. Mathews, *Idaho State University*
Peter C. Melchers, *University of Cologne (Germany)*
Claire Meljac, *Unité de Psychopathologie de l'Adolescent (France)*
Soo-Back Moon, *Catholic University of Hyosung (Korea)*
Arnold J. Moore, *Emporia State University*
Thomas D. Oakland, *University of Florida*
William Watson Purkey, *The University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Georgia Southern University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Valdosta State University*
Scott W. Snyder, *University of Alabama at Birmingham*
Bruce Thompson, *Texas A & M University*

GRADUATE STUDENT EDITORIAL BOARD
Margery E. Arnold, *Texas A & M University*
Vicki Benson, *The University of Alabama*
Alan Brue, *University of Florida*
Sue E. Castleberry, *Arkansas State University*
James Ernest, *University of Alabama at Birmingham*
Robin A. Groves, *Auburn University*
Harrison D. Kane, *University of Florida*
James C. Kaufman, *Yale University*
Sadegh Nashat, *Unité de Psychopathologie de l'Adolescent (France)*
Michael D. Scraper, *Kansas Newman College*
Sherry Vidal, *Texas A & M University*

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 1-8

Are Students Overly Confident in Their Mathematical Errors?

Talia Ben-Zeev
Yale University

Students often create erroneous algorithms for solving unfamiliar mathematics problems. Because these algorithms are rule-based, rather than being random, they result in solutions termed rational errors. Do students believe in the validity of their rational errors, or, rather, are students acting out of lack of alternatives? In order to investigate these questions, a recently developed methodology was used. Participants were taught how to perform addition in NewAbacus, a new number system. They were then tested on NewAbacus addition problems that were either familiar or new. Finally, participants were asked to provide ratings on the degree to which they believed that their solutions were accurate. Results indicated that participants believed in their errors more than was realistically warranted. An explanation for students' overconfidence is embedded in a more general discussion of the rationality underlying erroneous thinking.

When students commit errors on mathematics problems they often do so as a result of actively constructing erroneous rules or algorithms (Ashlock, 1976; Brown & VanLehn, 1980; Cox, 1975; Lankford, 1972; VanLehn, 1983; Young & O'Shea, 1981). Because these erroneous algorithms are rule-based and are consistent within an internal logic, they result in solutions termed *rational errors* (Ben-Zeev, 1995, 1996). For instance, VanLehn (1983, 1990) has found that students often commit what is known as the Smaller-from-Larger "bug:"

$$\begin{array}{r} 14 \\ \underline{-9} \\ 15 \end{array}$$

Students who commit this error subtract the top digit (e.g., "4") from the bottom digit (e.g., "9") instead of performing a borrowing action. This error may be considered as "rational" because it is consistent within the student's existing knowledge frame. That is, the student has learned that in single-digit subtraction one always subtracts the smaller from the larger digit (negative numbers are only introduced later on in the curriculum). The student, therefore, applies a logical approach by using an existing rule that has worked in past problem-solving episodes.

An important question is whether students believe in the validity of such errors, or are they aware that they are merely manipulating symbols in order to make a new and "foreign" problem look as familiar as possible? In research on clinical judgment, estimation of probability, and decision making (see for example, Einhorn & Hogarth, 1978; Fischhoff, Slovic, & Lichtenstein, 1977; Gigerenzer, Hoffrage, & Kleinbolting, 1991), researchers have demonstrated that people do tend to overestimate the degree of the accuracy of their judgments. Einhorn and Hogarth (1978) suggest that people's overconfidence can be explained by the fact that people tend to focus on the instances when their judgments are indeed correct, and either disregard, or do not have access to the instances when their judgments are false.

Can a similar "story" be told in the mathematical learning domain? Do students demonstrate overconfidence in their mistaken mathematical solutions? This paper will explore whether students actually have faith in their erroneous solutions or whether they are forced into providing answers to unfamiliar problems (e.g., leaving an answer blank is an unacceptable practice in school) and may simply lack better alternatives.

The overconfidence hypothesis has been underemphasized in research on errors in the mathematical domain, in particular, and in research on learning and skill acquisition, in general. An exception is work done by Payne and Squibb (1990). Payne and Squibb investigated rational errors, or "mal-rules," in students' performance on algebraic manipulation problems. Their examination of students' confidence was done as a subsidiary analysis for investigating the question of whether one could categorize rational errors into those that the majority of students believe in versus those that students do not. Payne and Squibb did not find support

---

The author wishes to thank Michel Ferrari, Henry Kaufman, and Robert Sternberg for invaluable comments on an earlier version of this paper. This work was supported by a graduate fellowship from Yale University. Correspondence should be sent to Talia Ben-Zeev, Department of Psychology, Yale University, P. O. Box 208205, New Haven, CT 06520-8205 or by e-mail to Talia@yalevm.cis.yale.edu.

for such a categorization scheme. They did find, however, that overall, students tended to believe that many of their erroneous solutions were accurate.

There are important theoretical and educational implications to a reality where students believe in their mathematical errors more than is realistically warranted. From a theoretical standpoint, demonstrating that students are overly confident in their rational errors, places research on school learning in the context of more general research on human cognition and problem solving, such as the work on judgment and decision making mentioned above. From an educational standpoint, students' beliefs in their erroneous solutions provide a challenge for educators who are trying to correct students' flawed representations. The latter point is especially important, because once a misconception is formed it may be extremely resistant to change (Resnick & Omanson, 1987; Rosnick & Clement, 1984).

The current study investigates the overconfidence hypothesis in the mathematical learning domain by using a recently developed methodology (Ben-Zeev, 1995). This methodology involved teaching participants a new number system, called NewAbacus, and observing participants' performance and perceived accuracy on addition problems in the new number system. In the Payne and Squibb (1990) experiment, participants may have exhibited overconfidence partially because they were already familiar with algebraic manipulation problems. Thus, the rationale for using the NewAbacus number system is that it controls for possible effects of familiarity on overconfidence. The question is whether, when the mathematical subject matter is truly new, students would still exhibit faith in their errors. Before discussing the specifics of past findings and current hypotheses, the reader is invited to become acquainted with number representation and addition in the NewAbacus number system.

The NewAbacus Number System

The NewAbacus number system can be seen in Figure 1. Each familiar base-10 digit is represented by two digits in NewAbacus. In the NewAbacus pair, the left digit is either 6 or 0, and the right digit ranges from 0 to 5. The sum of left and right digits in the NewAbacus pair, produces the familiar base-10 digit. For example, 7 in base-10 is equivalent to 61 in NewAbacus (6 + 1 = 7). Although 64 and 65 in NewAbacus sum up to be 10 and 11 in base-10 respectively, they are illicit in the NewAbacus system, because 64 and 65 violate the rule that each base-10 digit must be represented by two digits in NewAbacus. The correct representation for 10 and 11 in NewAbacus is 0100 and 0101 respectively.

0 = 00
1 = 01
2 = 02
3 = 03
4 = 04
5 = 05
6 = 60
7 = 61
8 = 62
9 = 63

10 = 0100
11 = 0101
12 = 0102
13 = 0103
14 = 0104
15 = 0105
16 = 0160
17 = 0161
18 = 0162
19 = 0163

20 = 0200
30 = 0300
40 = 0400
50 = 0500
60 = 6000
70 = 6100
80 = 6200
90 = 6300
100 = 010000

Figure 1. A list of base-10 numbers and their NewAbacus equivalents.

Addition in NewAbacus

The NewAbacus addition algorithm is divided into four main parts. They are, *no carry*, *carry into the 6 digit*, *carry from the 6 digit* and *carry into and from the 6 digit* (see Table 1). In the *no-carry* example, there is no difference between the base-10 and the NewAbacus addition algorithms. In the *carry into the 6 digit* example, adding column by column produces an intermediate solution where the right digit in a pair is equal to or greater than 6. In order to correct this violation, one must carry the 6 to the left and leave the remainder. For example, when the right column in the intermediate solution is 9, one carries a 6 and leaves a remainder of 3.

In the *carry from the 6 digit* example, a carry of a 1 is required *between*, rather than *within* a NewAbacus pair. When two 6s are added in one column the sum is 12. Therefore, one must carry a 1 to the next pair, and leave a remainder of 2. However, because the 2 remains in the left-digit, it violates the left-digit rule (it can only be 6 or 0). In order to correct the violation, one must sum the 2 with the right digit, to form a valid NewAbacus pair. Finally, the *carry into and from the 6 digit* example is a combination of the latter cases.

Table 1
Examples of the Different Parts of the Newabacus Addition Algorithm

Example 1 No carry		Example 2 Carry into 6		Example 3 Carry from 6		Example 4 Carry into and from 6	
		6		1		1 6	
6202		04		0260		62	
+ 0102		+05		+0161		+0405	
6304		~~9~~	invalid number	0421	sum columns	~~7~~	invalid number
		63	carry a six leave		carry a ten		so carry a six and
			remainder	04~~21~~	add digits		leave remainder
				0403	form valid	0521	sum columns
					number		carry a ten
						05~~21~~	add digits
						0503	form valid
							number

Students' performance on NewAbacus addition: What can it tell us about students' confidence in the validity of their solutions?

In a previous study (Ben-Zeev, 1995), participants received instruction on the NewAbacus number representation. Then, they received a set of worked-out examples of a particular part of the NewAbacus addition algorithm. Finally, participants were tested on a range of addition problems in NewAbacus where some problems were of a familiar type and most were new. In order to assess participants' confidence in their solutions, participants were asked to estimate how well they did (from 0-100%) on the addition test.

The present study provides a new analysis of students' overconfidence that had not been conducted in the 1995 study. Specifically, the current analysis predicts that if indeed participants believe that their errors are accurate, they will exhibit overconfidence. It is hypothesized, therefore, that participants will significantly overestimate their scores on the test. This prediction is aimed at showing that because students' mathematical errors are primarily systematic, deliberate, and therefore rational, students are not aware of the extent to which their performance is erroneous. That is, students may actually have faith in the validity of their mathematical errors.

Method

Participants

Participants were 80 Yale undergraduates consisting of thirty-two males and forty-eight females who either received a reward of $5 per session, or credit for an introductory psychology class. The race distribution of participants was as follows: 49 were Caucasian, 16 were African-American, ten were Asian, and five were Hispanic.

Procedure

Participants were given a list of NewAbacus numbers and their base-10 equivalents (see Figure 1), along with an explanation of the number representation rules. Next, participants were tested on their understanding of the NewAbacus number representation. The test was used for assessing participants' understanding, as well as for giving the participants an opportunity to practice with the NewAbacus numbers. After completing the number-representation test, participants were given worked-out examples of a specific type of NewAbacus addition problem (example types 1 through 4, as illustrated in Table 1). Participants were only allowed to advance to the next phase of the experiment if they had adequately understood the examples. Their understanding was assessed by asking participants verbally to simulate the steps in the worked-out examples. Next, participants were asked to solve 20 multicolumn addition problems that comprised the problem-set. Fifteen of these problems were new, and five were familiar.

Finally, after participants completed the problem set, they were asked to complete a questionnaire that asked them for information on their gender, major, SAT math score, number of college mathematics courses taken, most advanced mathematics courses taken, and grades received for each mathematics course taken. Participants were also asked to rate on a 1 to 7 Likert-type scale how comfortable they were with mathematics (where 1 indicates the highest comfort and 7 the lowest comfort), to rate their mathematical ability (where 1 indicates a very high ability and 7 a very low ability), and to express the degree of anxiety they had felt during the experiment (where 1 indicates the highest anxiety and 7 the lowest anxiety). Importantly, in order to assess confidence, participants were asked to estimate the percentage of problems (from 0% to 100%) they had accurately solved on the test.

Results

In order to determine if participants could accurately predict their scores on the NewAbacus addition test, a regression with perceived accuracy as the predictor, and actual accuracy as the dependent variable was performed. The regression accounted for 31% of the variance (using adjusted R^2), F (1, 78) = 36.36, $p < .0001$. Thus, in general, participants were fairly accurate at predicting whether they received a high or a low score. However, as can be seen in the scatter plot of the perceived by actual accuracy data (see Figure 2), the regression line, as well as most of the data points fall below the $y = x$ line. This result signifies a more subtle and significant pattern where the majority of participants tended to overestimate their scores. The data points for 59 participants fall below the $y = x$ line, whereas those for the other 21 participants are either on or above the regression line, $\chi^2(1, N = 80) = 9.03$, $p < .005$. When difference scores were computed for each participant (i.e., the difference between perceived accuracy and real accuracy), it was found that on average participants overestimated their score by 15%. When participants were divided into low-, medium-, and high-achieving groups (i.e., participants with actual scores between 0-33%, 34-66%, and 67-100% respectively), it was found that the high achievers correctly perceived their scores, whereas the others consistently overestimated their scores (see Table 2 for means and standard deviations).

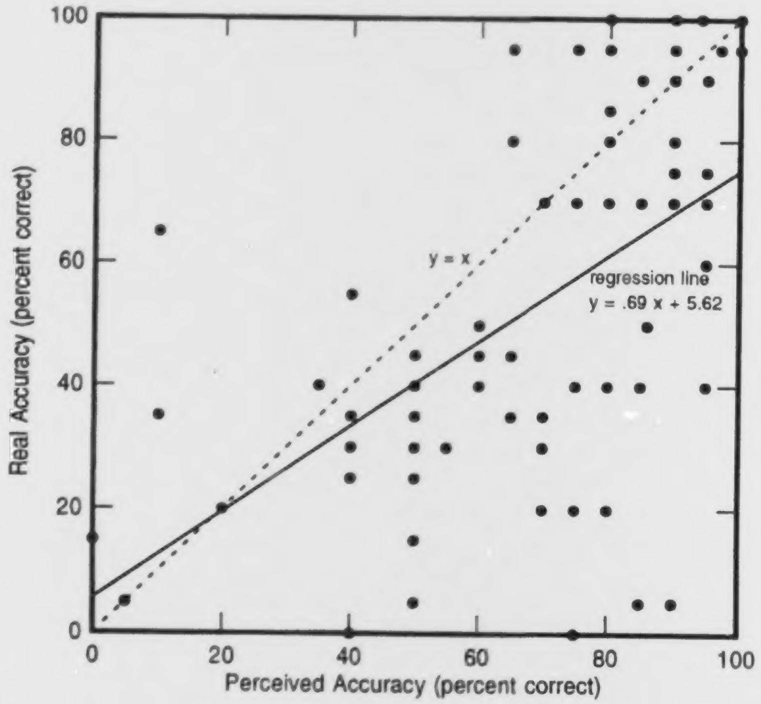


Figure 2. Participants' perceived versus actual accuracy on the addition test.

Table 2
The Low-, Medium-, and High-Achieving Groups' Actual and Difference Scores

Accuracy Level	*N*	Actual Score	Difference Score
Low achievers (0-33%)	21		
M		17.86	34.52
SD		10.56	27.79
Medium achievers (34-66%)	26		
M		43.30	17.35
SD		7.90	24.58
High achievers (67-100%)	33		
M		84.24	1.70
SD		11.73	14.19

Please note that difference scores were computed for each participant by subtracting perceived accuracy from real accuracy.

A one way analysis of variance (ANOVA) shows that the differences between the low-, medium-, and high-achieving groups' difference scores were highly significant, $F(2, 77) = 14.55$, $p < .0001$. A post-hoc contrast confirms the linear trend, $F(1, 77) = 28.79$, $p < .0001$. It demonstrates that participants' difference scores consistently decreased from the low through the medium to the high achievers. Thus, as participants performed less accurately and produced more erroneous algorithms, they tended to overestimate their scores. The overconfidence result is therefore robust, particularly given the fact that high-achieving participants reached a ceiling effect. That is, there was a cap on how much they could have overestimated their score (participants could not have overestimated their score above 100%), but not on how much they could have underestimated their scores.

The effects of the following individual differences on participants' actual accuracy on the test were tested: number of math courses taken, grade average on mathematics courses taken, degree of being comfortable with mathematics, number of mathematics classes taken in college, anxiety experienced during the experiment, SAT-quantitative score, gender, and race. A simultaneous multiple-regression analysis with the individual differences and perceived accuracy as the predictor variables and actual accuracy as the dependent variable accounted for 29% of the variance (using adjusted R^2), $F(8, 71) = 5.04$, $p < .0001$. Only perceived accuracy had a significant effect as a predictor variable (see Table 3).

Table 3
Perceived Accuracy and Individual Differences as Predictors of Actual Accuracy

Variable	Regression Scores		
	Parameter Estimate	*t*	α level
Perceived accuracy	0.61	4.31	0.0001
Sex	4.63	0.77	0.44
Race	-1.71	-0.29	0.77
Anxiety during experiment	0.55	0.27	0.79
Degree of comfort with math	-0.95	-0.44	0.65
SAT quantitative score	0.06	1.16	0.24
Number of math courses taken	2.05	0.71	0.47
Grade average on math courses	-2.36	-0.47	0.64

The same model but without perceived accuracy still resulted in a significant model but accounted for only 12% of the variance (using adjusted R^2), $F(7, 72) = 2.5$, $p < .05$. The only variable that emerged as significant was the anxiety participants experienced during the experiment, $t = 2.3$, $p < .05$. A third model was run without perceived accuracy or anxiety and was found to be nonsignificant, $F(6, 73) = 1.8$, $p > .09$. Thus, the strongest variable that predicted actual accuracy was the student's perceived accuracy. See Table 4 for the means and standard deviations of perceived accuracy and the individual differences.

Table 4
Participants' Perceived Accuracy and Individual Differences Scores

Individual Difference	Score	
	M	*SD*
Perceived accuracy	69	23.9
Anxiety during experiment	4.45	1.72
Degree of comfort with math	3.21	1.61
SAT quantitative score	693	69
Number of math courses taken	0.90	1.03
Grade average on math courses	3.52	0.63

An analysis of zero-order correlations revealed that perceived accuracy and anxiety had the strongest positive correlation with real accuracy. However, degree of comfortability with mathematics and the SAT quantitative score were also correlated modestly with real accuracy. See Table 5 for the zero-order correlations between all variables.

BEST COPY AVAILABLE

Table 5

Zero-order Correlations Between All Variables

Variable	1	2	3	4	5	6	7	8	9
1. Sex	----								
2. Race	-0.18	----							
3. Number of math courses	0.01	-0.02	----						
4. Comfort with math	-0.19	0.10	**-0.26**	----					
5. Perceived accuracy	0.15	-0.09	0.05	**-0.29**	----				
6. SAT Quantitative	-0.17	-0.10	**0.22**	**-0.41**	0.20	----			
7. Grade average	**0.24**	-0.04	-0.01	**-0.29**	0.09		----		
8. Anxiety	0.09	-0.10	0.09	**-0.41**	**0.53**	0.19	0.11	----	
9. Real accuracy	0.18	-0.11	0.14	**-0.28**	**0.56**	**0.26**	0.08	**0.35**	----

Please note that correlations in bold font are significantly different from chance. For correlations that have the absolute value of 0.22 and higher, $p < .05$. For correlations that have the absolute value of 0.28 and higher, $p < .01$. For correlations that have the absolute value of 0.35 and higher, $p < .001$.

Errors in NewAbacus resulted for the most part from a process of mis-induction. For instance, students who learned the *carry into the 6 digit* example commonly created the error *leave-a-6-carry-a-6-as-a-ten* (Ben-Zeev, 1995a):

$$\begin{array}{r} \mathbf{6} \\ 030063 \\ +\ \underline{020260} \\ 05\mathbf{6}263 \end{array}$$

This error "makes sense" because the student turned the familiar rule "IF a right digit in a pair is equal to or larger than 6, THEN carry the 6 and leave a remainder" into "IF a right or left digit in a pair is equal to or larger than 6, THEN carry the 6 and leave a remainder."

Another example of mis-induction comes from students who learned the *carry from the 6 digit* example. These students often committed errors such as *insert-pair-sum*:

$$\begin{array}{r} 010263 \\ +\ \underline{040302} \\ 0505\mathbf{65} \\ 0505\mathbf{0101} \end{array}$$

Students realized that the intermediate "65" was illicit, so they added the left and right digits and inserted the sum, "0101", directly into the solution. Students behavior was logically consistent because they turned the familiar rule "IF a pair of NewAbacus numbers is illicit and their sum is equal to or less then 6, THEN sum the digits and create a valid NewAbacus pair" into "IF a pair of NewAbacus numbers is illicit, THEN sum the digits and create a valid NewAbacus pair." For more examples of errors see Ben-Zeev (1995).

Discussion

Do students exhibit overconfidence in their mathematical errors? The answer is resoundingly positive. When students are asked to evaluate their performance on a new mathematical task, they tend consistently to overestimate its accuracy. The fact that students believe in their errors more than is realistically warranted lends support to the idea that students' errors result from a systematic approach to solving new problems.

A caveat is in order, however. This study assessed overconfidence in a population of Yale undergraduates who may be a particularly confident group of individuals. I believe, however, that the results of the study would still generalize to other populations for the following reasons: (a) There was a correlation between participants' perceived accuracy and their actual accuracy scores, and not just an "across-the-board" confidence effect where all participants expressed a high degree of belief in the validity of their solutions; and (b) The task itself was hard and thus presented a challenge even to Yale undergraduates.

The finding that students tend to be overconfident in their performance agrees with findings in the domains of judgment and decision making, which suggest that overconfidence may stem from the fact that people either do not have access to or choose to disregard information that seriously weakens or even eradicates the logical basis for their judgments. That is, people do not pay attention to disconfirming evidence (Einhorn & Hogarth, 1978).

Similarly, a robust finding in the mathematical problem-solving field is that worked-out examples have an important impact on error production, in that students tend to overgeneralize from familiar examples in order to solve new problems (Ben-Zeev, 1995, 1996; VanLehn, 1986, 1990). The danger of overly-relying on examples, however, is that examples often do not contain disconfirmatory instances, and thereby may facilitate the production of rational errors. For example, VanLehn (1986) has demonstrated that when students are given examples of only two-column subtraction problems, some students conclude that decrementing can only occur in the left-most digit.

A possible way to prevent students from becoming overconfident may be to provide them with "non-examples" of a particular mathematical concept or algorithm (see Shaughnessy, 1985). For instance, it is important that students be exposed to subtraction problems with a varied number of columns in order to avoid the subtraction error described above. Furthermore, it is important that "non-examples" be presented at the very beginning of skill acquisition because misconceptions have been shown to be extremely resilient to change once they are formed (Resnick & Omanson, 1987; Rosnick & Clement, 1984).

Another remedial approach may be to teach mathematical thinking as sense-making activity (Brown & Walter, 1993; Davis & Vinner; 1986; Greeno, 1983; Resnick & Omanson, 1987; Schoenfeld, 1991) by creating classroom environments that teach students to question the meaningfulness of their actions. For instance, students can be trained to ask What-If-Not questions (Brown & Walter, 1983; Brown & Walter, 1993) such as "what if there were more columns in a subtraction problem?," in order to challenge their assumptions and further their understanding.

Overconfidence does not have to be a hindrance to the mathematical learning process. It can also be a constructive means towards achieving a deeper understanding of one's reasoning. That is, by applying a metacognitive strategy of questioning one's confidence in a particular solution, the student may reach a deeper comprehension of his or her own mathematical thinking. The idea is not to cause students to self-doubt excessively, but to capitalize on the fact that their performance is both rational and fallible, such that they can learn from their mistakes. In fact, by committing mistakes and understanding their origin, students may achieve a stronger grasp of the relevant mathematical concepts and procedures, had they never committed errors at all (see also Smith, diSessa, & Roschelle, 1993).

Because of the theoretical and educational importance of students' overconfidence in their mathematical errors, further research is needed to more fully explore the nature and origin of this phenomenon. In the current study, the analysis of confidence levels was done by asking participants to provide ratings on their overall performance rather than on individual problems. Thus, part of the participants' confidence in their rational errors may have been the result of making careless errors or "slips" rather than committing truly algorithmic errors. Although a general analysis of confidence analysis is valuable it may therefore be slightly inflated. In future research, the new number system will be used to explore overconfidence on a problem by problem basis (see Payne & Squibb, 1990).

Now that students' overconfidence in their mathematical errors has been supported, further effort is needed to explore a variety of questions such as: Is overconfidence in mathematical errors a result of failing to attend to disconfirming information? How can we reduce overconfidence? Is there a gender difference in degree of overconfidence? Researching students' beliefs in their misconceptions in mathematics as well as in other learning domains may lead towards a better understanding of students' thinking and problem solving in general. In order to correct students' erroneous algorithms and beliefs we should aim to achieve a better understanding of how students create them in the first place.

References

Ashlock, R. B. (1976). *Error patterns in computation.* Columbus, Ohio: Bell and Howell.

Ben-Zeev, T. (1996). When erroneous mathematical thinking is just as "correct:" The oxymoron of rational errors. In R. J. Sternberg, & T. Ben-Zeev, (Eds.), *The nature of mathematical thinking* (pp. 55-79). Mahwah, NJ: Lawrence Erlbaum Associates.

Ben-Zeev, T. (1995). The nature and origin of rational errors in arithmetic thinking: Induction from examples and prior knowledge. *Cognitive Science, 19,* 341-376.

Brown, J. S., & VanLehn, K. (1980). Repair Theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4,* 379-426.

Brown, S. I., & Walter, M. I. (1983). *The art of problem-posing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, S. I., & Walter, M. I. (1993). Problem posing in mathematics education. In S. I. Brown, & M. I. Walter, (Eds.), *Problem posing: Reflections and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cox, L. S. (1975). Diagnosing and remediating systematic errors in addition and subtraction computation. *The Arithmetic Teacher, 22*, 151-157.

Davis, R. B., & Vinner, S. (1986). The notion of limit: Some seemingly unavoidable misconception stages. *Journal of Mathematical Behavior, 5*, 281-303.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*, 395-416.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 552-564.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Greeno, J. G. (1983). Forms of understanding in mathematical problem solving. In S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Learning and motivation in the classroom* (pp. 83-11). Mahwah, NJ: Lawrence Erlbaum Associates.

Lankford, F. G. (1972). *Some computational strategies of seventh grade pupils*. Charlottesville, VA: University of Virginia. ERIC Document No. ED 069 496.

Payne, S, & Squibb, H. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science, 14*, 445-481.

Resnick, L. B., & Omanson, S. F. (1987). Learning to understand arithmetic. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 3, pp. 41-95). Mahwah, NJ: Lawrence Erlbaum Associates.

Rosnick, P. & Clement, J. (1984). Learning without understanding: The effect of tutoring strategies on algebra misconceptions. *Journal of Mathematical Behavior, 3*, 3-27.

Schoenfeld, A. H. (1991). On mathematics as sense making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. F. Voss, D. N. Perkins, & J. W. Segal, (Eds.), *Informal reasoning and education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shaughnessy, M. J. (1985). Problem-solving derailers: The influence of misconceptions on problem-solving performance. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 399-415). Mahwah, NJ: Lawrence Erlbaum Associates.

Smith, J. P., diSessa, A. A., & Roschelle J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*, 115-163.

VanLehn, K. (1983). On the representation of procedures in repair theory. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 201-252). Mahwah, NJ: Lawrence Erlbaum Associates.

VanLehn, K. (1986). Arithmetic procedures are induced from examples. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 133-179). Mahwah, NJ: Lawrence Erlbaum Associates.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

Young, R. M., & O'Shea, T. (1981). Errors in students' subtraction. *Cognitive Science, 5*, 153-177.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 9-14

Effects of Learning Style Accommodation on Achievement of Second Graders

Carol Bugg Knight, Gerald Halpin, and Glennelle Halpin
Auburn University

The purpose of this study was to devise, to implement, and to evaluate an instructional model accommodating students' learning styles in the following areas: sound, light, temperature, design, and mobility. Tests of experimental and control group differences on grades in reading, mathematics, and language earned during treatment yielded a significant MANOVA. The follow-up ANOVAs showed that the groups differed significantly on grades in mathematics and language but not in reading. The students in the control group received significantly higher grades in mathematics and language. These results raise serious questions about accommodating individual differences in learning styles using the instructional model employed in this study.

Society's perception of education has changed dramatically in the past 50 years. Established practices of teaching and scheduling made the American educational system both successful and relatively well-respected during the first half of the century. Thus, schools and teachers were revered by the community, and any student who failed to learn did so because he or she failed to try with much diligence. Therefore, the blame fell on the student, not the teacher. Today, however, low achievement is directly blamed on the schools, teachers, and instructional programs (Dunn & Dunn, 1978).

According to Warner (1982), although some redundant and unnecessary experimentation was conducted in the 1970s, two important distinguishing characteristics of human learning have emerged: the diagnosis and treatment of individual learner differences and the management of the learning environment. Gregorc (1982) made a similar statement:

> education is making an insufficient impact on the human potential for learning. A primary contribution factor is our fragmented view of instruction . . . (and) learning style research has the potential to serve as a framework that could put the various facts and psychologies into a perspective and thereby lead us toward becoming a unified profession. (p. 31)

Keefe (1982) asserted that as society changes and costs increase, research on learning styles is becoming increasingly significant and that the key to effective schooling lies in understanding the range of student learning styles and designing appropriate materials that respond directly to the needs of the individual learner.

Ross (1980) pointed to the emerging understanding of how students with different styles function in various teaching-learning situations to form positive attitudes toward learning and promote a sense of satisfaction.

Learning styles can be defined as the "range of instructional strategies through which students typically pursue the act of learning" (Smith & Renzulli, 1984, p. 45). Because "most children can master the same content, how they master it is determined by their individual styles" (Dunn, Beaudry, & Klavas, 1989, p. 55).

Research by Dunn and Dunn (l978) yields at least 18 categories that, when classified, suggest that learners are affected by their immediate environment, their own emotionality, their sociological needs, and their physical needs. Further research corroborates that children learn more material more easily and retain more when they are taught through their preferred learning styles (Cafferty, 1980; Carbo, 1980; Domino, 1970; Douglass, 1979; Farr, 1971; Krimsky, 1982; Pizzo, l98l; Shea, 1983; Tannebaum, 1982; Urbschat, 1977; White, 1981).

Earlier research supports the accommodation of learning styles within the classroom. James (1962) and Pascal (1971), for example, provided support for a student-based instructional approach, pointing to the findings that educational outcomes are enhanced by giving students the opportunity to evaluate their learning style preferences and by allowing them to learn in their preferred modes of instruction. From the perspective of learning style, one major educational objective is to teach

---

Carol Bugg Knight is Vocational Counselor, Randolph/Roanoke Area Vocational School; both Gerald Halpin and Glennelle Halpin are Professors, Department of Educational Foundations, Leadership, and Technology at Auburn University. Please direct all correspondence to Gerald Halpin, 4036 Haley Center, Auburn University, AL 36849 (334) 844-3070, FAX: (334) 844-3072, Email: HALPIGE@mail.auburn.edu.

students how to learn and to manage their selection and use of various learning style strategies (Derry & Murphy, 1986). The accommodation of learning styles within the classroom can be a basis for providing instruction responsive to the needs of those being served. This clientele, of course, includes an increasingly diverse student population (Dunn et al., 1989).

Nowhere is this point more well expressed than by Mackinnon (1978):

> The wide range of individual differences surely must mean that there is no single method for nurturing creativity; ideally, the experiences we provide should be tailor-made, if not for individual students, at least for different types of students. We should remember that the same fire that melts the butter hardens the egg. (p. 171)

Teachers are relatively receptive to the concept of learning style accommodation within their classrooms because student achievement can be improved (Dunn & Dunn, 1987). As Jeter and Chauvin (1982) noted:

> Educators are keenly aware that each student possesses unique needs, interests, and abilities, and that each child should have an opportunity to pursue an effective instructional program at a pace that is challenging and interesting. (p. 2)

Administrators, on the other hand, are receptive to innovation only if proposed change does not prove too costly in terms of funding, special training, and equipment.

Thus, the purpose of this study was to devise a practical, easily implemented, relatively inexpensive learning style model that accommodated students' learning styles in the following areas: sound, light, temperature, design, and mobility. More specifically, the purpose of this investigation was to determine if grades earned in reading, mathematics, and language by students in an experimental group where learning environmental accommodations were made in areas of sound, light, temperature, design, and mobility differed significantly from those grades for students in a control group where no such accommodations were made. A further objective was to determine if the experimental manipulation was differen-tially effective for boys and girls and for black and white students.

According to Dunn (1987), research supports the importance of complementing the individual's learning style preferences with congruent instructional environments relative to sound, light, temperature, design, mobility, and time-of-day. The previously cited research studies and theoretical formulations would cause one to hypothesize that altering the environment would have a positive effect upon students' grades in reading, mathematics, and language.

Method

Subjects for this quasiexperimental investigation were 158 second graders in a county school system. Within each school were at least two sections of second grade, and students had been assigned at the beginning of the school year to sections on a random basis with an equal distribution of students in each section. Within each school the sections were randomly assigned to experimental and control groups with the assistance of the school's principal. The control group consisted of 34 boys and 35 girls/18 blacks and 51 whites. The experimental group contained 49 boys and 40 girls/31 blacks and 58 whites.

The students in the experimental groups were administered the Learning Style Inventory-Primary Version (Perrin, 1983) by a counselor within the school system. The principal investigator scored the inventories and derived a list for each teacher indicating preference for the areas of sound, light, temperature, design, and mobility of the students listed on the teacher's class roll. (Test-retest reliability coefficients reported by Perrin were .50, .74, .81, .88, and .82 for these respective areas.) Time-of-day was not considered because of scheduling conflicts which would have arisen.

Teachers of the experimental groups were then instructed to allow students to move to preferred areas of the room while doing seatwork in reading, mathematics, and language. Students preferring sound were fitted with headphones; students preferring warmth were seated in warmer areas of the room; students preferring bright light were seated in brightly lit areas of the room; and students preferring mobility were allowed to take brief walks in the classroom while completing seatwork. Conversely, students preferring no sound were seated in quiet sections of the room; students preferring cooler environments were seated in cool areas of the room; students preferring dim light were seated in areas of the room where less light was emitted from the fixtures; and students who did not express mobility as a preference were allowed to sit for long periods of time.

Measures of light and temperature were made more accurate with the use of thermometers and the measurement of footcandles emitted by the light fixtures in certain portions of the classroom. Teachers were informed of these designated areas, and they then placed students according to student preferences while students did their assigned seatwork in reading, mathematics, and language. Students and teachers in the control group were provided no special instructions, but the teachers were informed that they were to continue teaching just as they usually did.

Learning style accommodation commenced at the beginning of the fourth 6-weeks grading period and continued through the fifth 6-weeks. The teacher-assigned grades in reading, mathematics, and language arts were recorded for each student in the experimental and control groups for each of the six grading periods of the school year. These were designated from an *A+* as *13* to an *F* as *1*. In addition to age, sex and race of each child were recorded as well as the school and group (experimental or control) to which the child belonged.

Multivariate and univariate analyses of variance were used in the analyses of the resulting data in this study with the individual student being the unit of analysis. In educational research, students are often subjected to treatment in a group setting, but the statistical analysis is conducted with the student as the unit of analysis. Such a practice may jeopardize the assumption of independence. However, should the class have been selected as the unit of analysis, the resulting statistical power would have been grossly inadequate. Further, should a modification have been made in the design so that each student would have been exposed to the treatment individually in an isolated setting, the external validity of the study would have been weakened.

Results

Even though students were randomly assigned to classes and classes were randomly assigned to experimental and control groups, an initial test was conducted to evaluate group equivalence. The results of the multivariate analyses, MANOVA, revealed no significant pretreatment differences between the groups on Stanford Achievement NCE scores on reading and math and grades in reading, mathematics, and language, $F(5, 125) = 1.26, p = .29$. The Wilks' lambda of .951 indicates less than 5% of the variance in the synthetic variable was attributed to pretreatment differences. Means and standard deviations are reported in Table 1.

Next, a 2 x 2 x 2 MANOVA was conducted to determine the effects of the treatment and the interactive effects of treatment condition and sex and/or race on the post-treatment grades earned in reading, mathematics, and language. In the overall MANOVA, Wilks' lambda was not found to be significant for any of the interaction combinations of treatment by sex by race as can be seen in Table 2. Adhering to the general guidelines for the MANOVA procedure regarding significance, further interpretation of the interactions at the univariate level was not appropriate.

Table 1
Pretreatment Means and Standard Deviations for Stanford Achievement Test Scores in Reading and Mathematics and for Grades in Reading, Mathematics, and Language for Experimental and Control Groups

Group	Mean	SD	Number
	Stanford Achievement Reading		
Experimental	58.18	19.10	57
Control	59.31	18.40	74
	Stanford Achievement Mathematics		
Experimental	64.90	21.44	57
Control	60.11	19.36	74
	Mathematics Grade		
Experimental	31.58	6.17	57
Control	32.07	7.10	74
	Language Grade		
Experimental	25.33	10.09	57
Control	26.68	10.44	74
	Reading Grade		
Experimental	23.19	10.46	57
Control	24.68	10.81	74

Table 2
Multivariate Analysis of Variance Effects of Interaction, Treatment, Sex, and Race on Post-Treatment Grades in Reading, Mathematics, and Language

Source	Wilks' lambda	Hypothesis *df*	Error *df*	*F*
Group (A)	.93	3	142	3.79*
Sex (B)	.85	3	142	8.48**
Race (C)	.96	3	142	2.16
A x B	.98	3	142	.98
A x C	.99	3	142	.18
B x C	.96	3	142	2.20
A x B x C	99	3	142	.79

* $p < .05$
** $p < .001$

There was a significant main effect for treatment in the MANOVA, $F(3, 142) = 3.79$, $p = .01$. Follow-up univariate analyses of variance (ANOVAs) for the three dependent variables are presented in Table 3.

Table 3
Factorial Analysis of Variance of Effects of Treatment, Sex, Race, and Their Interactive Effects on Grades in Reading, Mathematics, and Language

Source	*df*	*SS*	*MS*	*F*	R^2
		Reading			
Group (A)	1	354.33	354.33	3.19	.02
Sex (B)	1	1343.57	1343.57	12.08***	.07
Race (c)	1	391.42	391.42	3.52	.02
A x B	1	222.31	222.31	2.00	.01
A x C	1	.02	.02	.00	.00
B x C	1	6.71	6.71	.06	.00
A x B x C	1	103.48	103.48	.93	.01
Error	144	16021.55	111.26		
Total	151	18402.40	121.87		
		Mathematics			
Group (A)	1	688.84	688.84	8.005**	.06
Sex (B)	1	1303.929	1303.92	15.29***	.11
Race (C)	1	559.55	559.55	6.56***	.05
A x B	1	159.31	159.31	1.87	.01
A x C	1	18.13	18.13	.21	.00
B x C	1	277.94	277.94	3.26	.02
A x B x C	1	75.18	75.18	.88	.01
Error	144	12280.11	85.28		
Total	151	14930.52	98.88		
		Language			
Group (A)	1	688.46	688.46	7.74**	.04
Sex (B)	1	2099.78	2099.78	23.61**	.13
Race (C)	1	322.117	322.11	3.62	.02
A x B	1	211.77	211.77	2.38	.01
A x C	1	5.51	5.51	.06	.00
B x C	1	78.00	78.00	.88	.00
A x B x C	1	52.32	52.32	.59	.00
Error	144	12805.44	88.93		
Total	151	16068.08	106.41		

* $p < .05$
** $p < .01$
*** $p < .001$.

Significant univariate *F* values were found for grades in mathematics [$F(1, 144) = 8.01$, $p = .01$] and for grades in language [$F(1, 144) = 7.74$, $p = .01$] but not for grades in reading [$F(1, 144) = 3.19$, $p = .08$]. Subjects in the control group scored significantly higher ($M = 27.31$) than subjects in the experimental group ($M = 23.43$) on grades earned in mathematics. Further, subjects in the control group scored significantly higher ($M = 27.34$) than subjects in the experimental group ($M = 23.49$) on grades in language. The control group mean ($M = 25.90$) was not significantly higher than the experimental group mean ($M = 23.15$) on grades in reading (see Table 4).

Table 4
Post-Treatment Means and Standard Deviations for Grades in Reading, Mathematics, and Language for Experimental and Control Groups

Group	Mean	SD	Number
	Mathematics Grade		
Experimental	23.43	10.36	65
Control	27.31	9.34	87
	Language Grade		
Experimental	23.49	10.92	65
Control	27.34	9.57	87
	Reading Grade		
Experimental	23.15	10.81	65
Control	25.90	11.13	87

Discussion

The results indicate that the experimental and control groups did not differ prior to the beginning of the treatment on grades earned in reading, mathematics, and language as would be expected under conditions where subjects were randomly assigned to groups and groups were randomly assigned treatment conditions. What was not expected based upon the literature reviewed was that the control group would have significantly higher grades in mathematics and language after treatment. What are some possible explanations for such an anomaly? One possible explanation is the accommodations of the learning environment involving moving students around for lighter/darker, cooler/warmer, noisier/quieter work areas were actually disruptive and negatively impacted student achievement of teacher classroom objectives in mathematics and language. It may be that had the accommodations been employed during instruction and for longer periods of time significant results favoring the intervention might have been found. Another possibility is that the Dunn and Dunn model is of questionable educational value.

Whatever the explanation, results of this study cast doubt on the benefit of such accommodations and suggest that educators should conduct further study before implementing interventions no matter how popular the package. While, no doubt, there are environments in which learning is maximized, it just may be that some of the most highly touted interventions may not be the most sound educationally. What happens in the classroom should be based on significant research. That more research is necessary before learning style accommodations be made is the major conclusion of this study. If the research does not support the learning style interventions, then they should not become classroom practice.

References

Cafferty, E. (1980). *An analysis of student performance based upon the degree of match between the educational cognitive style of the teachers and the educational cognitive style of the students.* Unpublished doctoral dissertation, University of Nebraska.

Carbo, M. (1980). An analysis of the relationship between the modality preferences of kindergartners and selected reading treatments as they affect learning of basic eight-word vocabulary (Doctoral dissertation, St. John's University, New York). *Dissertation Abstracts International, 41,* 1389A.

Derry, J. J., & Murphy, D. A. (1986). Designing systems that train learning ability: From theory to practice. *Review of Educational Research, 53,* 243-251.

Domino, G. (1970). Interactive effects of achievement orientation and teaching style on academic achievement. *ACT Research Report* 39: 1-9.

Douglass, C. B. (1979). Making biology easier to understand. *American Biology Teacher, 41,* 277-299.

Dunn, K., & Dunn, R. (1987). Dispelling outmoded beliefs about student learning. *Educational Leadership, 44,* 55-61.

Dunn, R. (1987). Research on instructional environments: Implications for student achievement and attitudes. *Professional School Psychology, 2*(1), 43-52.

Dunn, R., Beaudry, J. S., & Klavas, A. (1989). Survey of research on learning styles. *Educational Research, 45,* 50-58.

Dunn, R., & Dunn K. (1978). *Teaching students through their individual learning styles: A practical approach.* Englewood Cliffs, NJ: Prentice Hall.

Farr, B. J. (1971). *Individual differences in learning: Predicting one's more effective learning modality.* Unpublished doctoral dissertation, Catholic University of America.

Gregorc, A. F. (1982). Learning style/brain research: Harbinger of an emerging psychology. In *Student learning styles and brain behavior: Programs, instrumentation, research.* Reston, VA: National Association of Secondary School Principals.

James, N. E. (1962). Personal preference for method as a factor in learning. *Journal of Educational Psychology, 53,* 43-47.

Jeter, J., & Chauvin, J. (1982). Individualized instruction: Implications for the gifted. *Roeper Review, 5,* 2-3.

Keefe, J. W. (1982). Assessing student learning styles: An overview. In *Student learning styles and brain behavior: Programs, instrumentation, research.* Reston, VA: National Association of Secondary School Principals.

Krimsky, J. S. (1982). *A comparative study of the effects of matching and mismatching fourth grade students with their learning style preferences for the environmental element of light and their subsequent reading speed and accuracy scores.* Unpublished doctoral dissertation, St. John's University, Jamaica, NY.

Mackinnon, D. (1978). *In search of human effectiveness: Identifying and developing creativity.* Buffalo, NY: Creative Education Foundation.

Perrin, J. (1983). *Learning style inventory, primary version.* Jamaica, NY: St. John's University, Learning Style Network.

Pascal, C. E. (1971). Instructional options, option preferences, and course outcomes. *The Alberta Journal of Educational Research, 17,* 1-11.

Pizzo, J. (1981). *An investigation of the relationship between selected acoustic environments and sound, an element of learning style as they affect sixth grade students' reading achievement and attitudes.* Unpublished doctoral dissertation, St. John's University, Jamaica, NY.

Ross, H. G. (1980). Matching achievement styles and instructional environments. *Contemporary Educational Psychology, 5,* 216-226.

Shea, T. (1983). *An investigation of the relationship(s) among preferences for the learning style element of design, selected instructional environments and reading achievement of ninth grade students to improved administrative determinations concerning effective educational facilities.* Unpublished doctoral dissertation, St. John's University, Jamaica, NY.

Smith, J. H., & Renzulli J. S. (1984). Learning style preferences: A practical approach for classroom teachers. *Theory into Practice, 23,* 44-50.

Tannebaum, R. (1982). *An investigation of the relationship(s) between selected instructional techniques identified field dependent and field independent cognitive styles as evidenced among high school students enrolled in studies in nutrition.*

Unpublished doctoral dissertation, St. John's University, Jamaica, NY.

Urbschat, K. S. (1977). *A study of preferred learning models and their relationship to the amount of recall of CVC trigrams*. Unpublished doctoral dissertation, Wayne State University.

Warner, W. (1982). Cognitive style mapping by the Hill Model. In *Student learning styles and brain behavior: Programs, instrumentation, and research*. Reston, VA: National Association of Secondary School Principals.

White, R. T. (1981). *An investigation of the relationship between selected instructional methods and selected elements of emotional learning style upon student achievement in seventh-grade social studies*. Unpublished doctoral dissertation, St. John's University, Jamaica, NY.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 15-21

Differences in Reading and Math Achievement Test Scores for Students Experiencing Academic Difficulty

John R. Slate
Valdosta State University

Craig H. Jones
Arkansas State University

Relationships between achievement scores in reading and mathematics on the KeyMath-Revised, Peabody Individual Achievement Test-Revised (PIAT-R), Wechsler Individual Achievement Test, and Woodcock Reading Mastery Test-Revised were investigated in a sample of 366 students experiencing academic difficulty. Both reading and mathematics scores were lower on the PIAT-R than were comparable scores on other tests. Scores on subtests purporting to measure the same reading or mathematics constructs were positive and statistically significant. These correlations, however, tended to be modest in size ranging from .23 to .61. Overall, reading subtests averaged 39% shared variance with subtests purporting to measure the same construct. Mathematics subtests averaged 25% shared variance with subtests purporting to measure the same construct. Implications of our findings for practitioners, researchers, and test publishers are discussed.

When a student experiences repeated academic failure in school, he or she will often be referred for psychological evaluation to determine eligibility for special education services (Fuchs & Fuchs, 1988). This psychological evaluation typically will involve consideration of scores from one or more measures of academic achievement, an individual intelligence test score, and other information pertinent to diagnosing disability under current special education rules and regulations. Although most examiners use the Wechsler Intelligence Scale for Children to assess intelligence (Piotrowski & Keller, 1989), the tests used to assess academic achievement vary widely from one school district to the next. Among the measures commonly used to assess academic achievement are the KeyMath-Revised (KM-R; Connolly,1988), the Peabody Individual Achievement Test-Revised (PIAT- R; Markwardt, 1989), the Wechsler Individual Achievement Test (WIAT; Wechsler, 1992), and the Woodcock Reading Mastery Tests-Revised (WRMT-R; Woodcock, 1987).

Some authors have defended the wide variation in the achievement tests used for diagnosis as being needed to provide flexibility in matching test content to particular school curricula (Shapiro & Derr, 1987; Webster & Braswell, 1991). When there is a mismatch between test content and curriculum, low test scores could reflect an artifact of the assessment process rather than low student achievement. The more freedom an examiner has in selecting an achievement test, the more likely the examiner can match test content to the curriculum. The wide variation in the achievement tests used to assess students, however, can also cause problems in diagnosis. Achievement tests purported to measure the same construct may not actually do so. In addition, even when the same construct is measured, differences in standardization procedures could result in students receiving different scores depending upon which achievement test is used (e.g., Caskey, 1985; Caskey, Hylton, Robinson, Taylor, & Washburn, 1983; Eaves, Darch, & Haynes, 1989; Shapiro & Derr, 1987).

The processes of test selection and score interpretation would be facilitated if examiners had access to data on the comparability of scores from the most commonly used achievement tests. Unfortunately, such data are not readily available. A CD-ROM search of ERIC and Psychological Abstracts failed to identify any published articles in which scores on the above-mentioned achievement tests were related to each other. A hand search of

---

John R. Slate is a Professor in the Department of Educational Leadership at Valdosta State University. Craig H. Jones is a Professor in the Department of Psychology and Counseling at Arkansas State University (email: cjones@kiowa.astate.edu). Please address correspondence regarding the paper to John R. Slate, Department of Educational Leadership, Valdosta State University, Valdosta, GA 31698-0090 or by-email to jslate@grits.valdosta.peachnet.edu.

The data were collected as part of a large project designed to explore multiple aspects of testing in special education diagnosis. Additional data regarding the students' IQs, and the relationship of their IQs to their achievement test scores, can be found in Slate (1994), Slate (1995a), Slate (1995b), and Slate (1995c). Data from several subpopulations were combined in the present study because separate analyses would have resulted in very low sample sizes for some statistical tests.

other sources, however, did locate two recent, unpublished studies in which relationships between scores on the KM-R, PIAT-R, WIAT, and WRMT-R were investigated. In one of these studies, Slate (1996) compared achievement test scores for students with learning disabilities and found that math and reading scores on the PIAT-R were lower than were comparable scores on the WIAT and the KM-R. In addition, reading scores on the WRMT-R were lower than were reading scores on the WIAT. In the other study, Slate and Saarnio (1996) compared achievement test scores for students with mental retardation and found that reading scores were lowest on the WRMT-R and highest on the WIAT. Reading scores on the PIAT-R fell in between scores on the WRMT-R and the WIAT. In addition, math scores on the PIAT-R were lower than were scores on the KM-R. Thus, the findings of Slate and Saarnio (in press) for students with mental retardation were comparable to those of Slate (1996) for students with learning disabilities. Achievement tests purporting to measure the same or similar constructs yielded significantly different mean scores. In addition, the researchers in both studies reported that the intercorrelations between these tests ranged from low to moderate. That is, the shared variance for reading comprehension subtests ranged from 35% to 37%, and the shared variance ranged from 19% to 21% for total math scores. Thus, the available evidence, albeit limited, indicates that the achievement test scores of students undergoing special education assessment will vary substantially depending upon the achievement test that is used.

Because of the importance of achievement test scores in determining whether or not students with learning difficulties qualify for special services, replication and extension of the results obtained by Slate (1996) and Slate and Saarnio (in press) are needed. That is, assessment specialists need to know if these data are both reliable and applicable to students other than those with learning disabilities or mental retardation. The present study was, therefore, conducted to determine the extent to which different standardized achievement tests produced similar scores in a sample of students who had been referred for a psychological evaluation but failed to qualify for special services.

Method

Data were collected from three educational cooperatives serving school districts in the Mississippi Delta region of northeast Arkansas. Achievement test scores were obtained on 366 White students (222 males, 144 females) with an average age of 9.8 years (*SD* = 2.9 years). [Note. Because data were present on fewer than 5 African-American students, their data were discarded from statistical analyses.] The mean WISC-III Full Scale IQ for these students was 81.2 (*SD* = 10.0), with a mean Verbal IQ of 82.0 (*SD* = 11.2) and a mean Performance IQ of 83.3 (*SD* = 10.1). All students on whom data were collected had been referred for psychological evaluations because of academic difficulties but had failed to meet eligibility criteria for special education under Arkansas rules and regulations (*Program Standards and Eligibility Criteria for Special Education*, 1987).

Test scores were recorded from students' special education folders and, therefore, reflect the test scores used in making the determination that they did not meet eligibility criteria for special education. The examiners employed by the three educational cooperatives involved in the study most commonly administered the KM-R, the PIAT-R, the WIAT, and the WRMT-R to assess students' academic achievement. Thus, scores on these standardized tests became the focus of the data analysis. Using data from actual special education evaluations has both advantages and disadvantages. The major disadvantages are (a) researchers are unable to select the specific tests that are the focus of the investigation but, rather, must select the tests to be analyzed based upon examiner preferences, and (b) scores for specific tests are not available for all students because examiners do not give every test to every student. Thus, in the present study, sample sizes for statistical comparisons ranged from 34 to 84 students depending upon the specific achievement tests involved.

On the other hand, using actual assessment data has the advantages of (a) focusing the investigation on the tests that are most likely to be used to make diagnoses rather than the tests of interest to the researchers, and (b) allowing the data to reflect the relationships between test scores that occur under actual testing conditions rather than under artificial conditions contrived by the researchers. Because we were interested in how achievement test scores affect actual special education decisions, we believed that the advantages of using natural data outweighed the disadvantages in the present study.

Although we were unable to select the specific tests that were the focus of the investigation, the four achievement tests most commonly used by the examiners were broadly representative of special education practices. That is, three of these tests, or previous versions of them, were among the tests Connelly (1985) found to be most frequently used by special education teachers. The exception is the WIAT which was published after Connelly's study. In addition, the assessment literature indicates that the achievement tests that were the focus of the present study have strong psychometric properties

(Salvia & Ysseldyke, 1995). Their use as either screening measures or as in-depth measures of academic achievement is also well supported by the literature (Beck, 1992; Benes, 1992; Bracken, 1988; Finley, 1992; Rogers, 1992; Salvia & Ysseldyke, 1995).

Results

Means, standard deviations, and sample sizes for all reading and mathematics subtest scores are presented in Table 1. Reflective of the fact that all of the students were experiencing academic difficulties, the mean scores for all subtests were in the 80's. A restricted range of scores was present as evidenced by standard deviations of less than 15 for all subtests.

Table 1
Means, Standard Deviations, and Sample Sizes for Each Achievement Test

Test Score	Mean	*SD*	*n*
PIAT-R			
Reading Recognition	85.9	10.1	176
Reading Comprehension	87.0	12.0	178
Total Reading	85.1	10.6	180
Math	86.7	12.6	176
KM-R			
Basic Concepts	87.4	11.4	207
Applications	87.3	11.8	203
Operations	85.1	12.3	207
Total Test	85.5	10.7	204
WRMT-R			
Basic Reading Skills	88.1	11.6	248
Reading Comprehension	83.8	9.7	242
Total Reading	86.7	10.3	235
WIAT			
Basic Reading Skills	89.3	13.0	192
Reading Comprehension	86.1	10.1	176
Math Reasoning	87.8	8.8	186
Numerical Operations	87.5	10.5	155

Statistical comparisons of all subtest scores purported to measure the same construct were made using *t*-tests for correlated measures. The results for reading subtests are presented in Table 2. [Note. The means and standard deviations reported in Tables 2 and 3 may differ from the means and standard deviations reported in Table 1 because only the test scores of those students who had been administered both tests analyzed were included in the calculation of the mean and standard deviation for each comparison.] Because seven statistical tests were conducted on the reading subtest data, the alpha level for statistical significance was set at .007 by dividing the conventional alpha level of .05 by the number of *t*-tests being conducted to reduce the possibility of a Type I error. Using this alpha level, three of the reading subtest comparisons yielded statistically significant differences between scores. All of the significant differences involved scores on the PIAT-R. The PIAT-R Reading Recognition subtest scores were 4.5 points lower than were scores on the WRMT-R Basic Reading and 5.5 points lower than were scores on the WIAT Basic Reading scores. In addition, PIAT-R Total Reading scores were 4.3 points lower than the WRMT-R Total Reading score. Thus, in all statistically significant comparisons, scores on the PIAT-R were lower than purportedly comparable reading scores on other achievement tests. Note that using an alpha of .01 would have also resulted in WRMT Reading Comprehension scores being lower than were WIAT Reading Comprehension scores.

The results for mathematics subtests are presented in Table 3. Because eight *t*-tests were calculated, an alpha level of .006 (i.e., .05 divided by 8) was used to reduce the likelihood of a Type I error. Using this alpha level, three of the eight *t*-tests yielded statistically significant differences. Again, all three differences involved students receiving lower scores on the PIAT-R than on other achievement tests. Specifically, the PIAT-R Math subtest yielded average scores 5.8, 4.9, and 4.0 points lower than the KM-R Basic Concepts, Applications and Total scores, respectively. Note that the PIAT-R Math subscores would have also been significantly different from KM-R Operations and WIAT Math Reasoning had an alpha of .05 been used.

The amount of shared variance between various subtest scores was determined by calculating the simple Pearson's *r* for each pair of reading subtest scores and each pair of mathematics subtest scores that are purported to measure the same construct. The correlations for reading subtests are presented in Table 4. All of these correlations were positive and statistically significant at the .001 level. The amount of shared variance ranged from a low of 20% between Reading Comprehension scores on the PIAT-R and the WIAT to a high of 61% between Basic Reading Skills scores on the WIAT and the WRMT-R. Overall, reading scores purported to measure the same construct averaged 39% common variance, with Basic Reading scores averaging 43% shared variance and Reading Comprehension scores averaging 31% shared variance.

Table 2
Means, Standard Deviations, and *t*-Tests for Reading Comparisons

Test Scores	*M*	*SD*	*t*	*df*	*p*
PIAT-R Reading Recognition	84.9	9.4			
WRMT Basic Reading	89.4	10.1	-5.59	100	.0001
PIAT-R Reading Recognition	84.9	9.8			
WIAT Basic Reading	90.4	8.2	-5.92	69	.0001
PIAT-R Reading Comprehension	84.8	12.1			
WRMT Reading Comprehension	84.4	10.0	0.45	98	> .05
PIAT-R Reading Comprehension	84.0	12.6			
WIAT Reading Comprehension	85.9	9.6	-1.31	65	> .05
PIAT-R Total Reading	83.4	10.4			
WRMT Total Reading	87.7	9.9	-5.65	98	.0001
WRMT Basic Reading	88.5	12.2			
WIAT Basic Reading	89.3	10.1	-1.34	151	> .05
WRMT Reading Comprehension	83.6	9.4			
WIAT Reading Comprehension	85.4	8.4	-2.48	139	.01

Note. Means and standard deviations may differ from the Table 1 means and standard deviations because only the data on those students who had both test scores used in the *t*-test analysis were calculated.

Table 3
Means, Standard Deviations, and *t*-Tests for Math Comparisons

Test Scores	*M*	*SD*	*t*	*df*	*p*
PIAT-R Math	82.7	12.7			
KM-R Basic Concepts	88.5	12.1	-3.74	67	.0001
PIAT-R Math	83.2	12.5			
KM-R Applications	88.1	11.7	-3.33	65	.001
PIAT-R Math	82.7	12.7			
KM-R Operations	86.0	10.9	-2.24	67	.029
PIAT-R Math	82.7	12.7			
KM-R Total	86.8	10.4	-3.22	67	.002
PIAT-R Math	83.6	13.5			
WIAT Math Reasoning	87.1	7.5	-2.22	69	.03
PIAT-R Math	85.3	13.9			
WIAT Numerical Operations	87.5	9.7	-1.14	60	> .05
KM-R Applications	88.5	11.3			
WIAT Math Reasoning	87.7	8.2	0.93	130	> .05
KM-R Operations	86.5	11.9			
WIAT Numerical Operations	87.8	10.8	-1.50	113	> .05

Note. Means and standard deviations may differ from the Table 1 means and standard deviations because only the data on those students who had both test scores used in the *t*-test analysis were calculated.

Table 4
Correlations of WIAT, WRMT-R, and PIAT-R Reading Subtests Purporting to Measure the Same Construct

Construct Test Scores	*r*	r^2	*n*
Basic Reading Skills			
WRMT-R with WIAT	.78	.61	152
WRMT-R with PIAT-R	.52	.27	99
WIAT with PIAT-R	.64	.41	70
Reading Comprehension			
WRMT-R with WIAT	.54	.29	140
WRMT-R with PIAT-R	.67	.45	99
WIAT with PIAT-R	.45	.20	66
Total Reading			
WRMT-R with PIAT-R	.72	.52	99

Note. All correlations are significant at the .001 level.

The correlations for math subtests are presented in Table 5. All of these correlations were positive and significant at the .001 level except for the correlation between the WIAT Numerical Operations and the PIAT-R Math subtests which was significant at the .05 level. The amount of shared variance ranged from a low of 5% between Numerical Operations scores on the PIAT-R and the WIAT to a high of 46% between Numerical Operations scores on the KM-R and the WIAT. Overall, the mathematics subtest scores averaged 25% common variance, with Math Reasoning subtests averaging 22% shared variance and Numerical Operations subtests averaging 24% shared variance.

Table 5
Correlations of WIAT, KM-R, and PIAT-R Mathematics Subtests Purporting to Measure the Same Construct

Construct Test Scores	*r*	r^2	*n*
Math Reasoning			
WIAT with KM-R	.51	.26	131
WIAT with PIAT-R	.36	.13	70
KM-R with PIAT-R	.52	.27	66
Numerical Operations			
WIAT with KM-R	.68	.46	114
WIAT with PIAT-R	.23	.05	61
KM-R with PIAT-R	.47	.22	68
Total Math			
KM-R with PIAT-R	.62	.38	68

Note. All correlations are significant at the .001 level except the correlation between the PIAT Math and the WIAT for Numerical Operations which was significant at the .05 level.

Discussion

The results of the present study substantially replicate the previous findings of Slate (1996) and Slate and Saarnio (1996). That is, four of the most commonly administered standardized achievement tests often yielded significantly different scores on subtests purported to measure the same construct. Significant differences were found on three of seven comparisons involving reading scores. All of these differences occurred because students received significantly lower scores on the PIAT-R than on other achievement tests. Similarly, three of eight comparisons involving mathematics subtest scores were significantly different. All three significant differences were again due to students receiving lower scores on the PIAT-R.

For readers who are aware of the Flynn effect, it does not explain the differences among achievement tests purporting to measure the same or similar constructs. That is, Flynn (1984; 1987) has provided convincing evidence that an average gain of three points in IQs occurs every decade in the United States. Therefore, IQ tests with normative dates a decade or decades apart will provide different scores, solely as a function of the publication date (Bracken, 1988). In this study, all tests had similar publication dates, ranging from 1987 to 1992, with the PIAT-R having the middle publication date, 1989, providing consistently lower scores than the other tests. Thus, factors other than Flynn's effects are the causes for the statistically significant achievement differences reported herein.

Even when statistically significant differences between test scores were not found, the amount of shared variance was typically modest. The largest amount of shared variance was 61% between Basic Reading Skills scores on the WIAT and the WRMT-R. Of 14 total correlations, eight indicated less than 30% shared variance with a low of 5% between the PIAT and the WIAT Numerical Operations scores. In general, reading scores were more strongly associated with each other than were mathematics scores. This finding is not surprising given that mathematics scores on standardized tests are often confounded by reading ability (Marston, 1989).

Although the results of Slate (1996), Slate and Saarnio (1996), and the present study all indicate that scores of various standardized achievement tests are modestly related to each other at best, there is an important difference in the specific relationships that were found. In both the Slate (1996) study and the present study, which involved the scores of students with learning disabilities and students who did not qualify for special education, respectively, scores on the PIAT-R were consistently lower than were scores on other achievement tests. In the Slate and Saarnio (1996) study, which involved the scores of students with mental retardation, PIAT-R mathematics scores were lowest on the PIAT-R but not reading scores. Among the students with mental retardation, reading scores were lowest on the WRMT-R. This finding suggests that the interrelations among achievement test scores may vary to some degree as a function of students' disabilities. Certainly, the PIAT-R needs to be examined inasmuch as its scores were consistently lower than scores provided by other tests.

The present study was conducted with a sample of students tested by examiners working for three educational service cooperatives in the same geographical region. Thus, the extent to which the present results can be generalized to other populations is an open question. Clearly, additional research is needed to delineate the relationships among scores on various standardized achievement tests with greater certainty. At very least, the present research should stimulate such research.

If we assume that the present results are generalizable, they indicate that standardized achievement tests are not routinely interchangeable. The achievement test selected for use by a psychological examiner can, therefore, profoundly affect whether or not a student receives special education services. For example, students suspected of a learning disability in mathematics are most likely to qualify for special services if tested with the PIAT-R because the lower scores characteristic of this test are most likely to produce the discrepancy between IQ and achievement needed for diagnosis. Thus, practitioners should carefully select achievement tests based upon the best match between test content and likely problem areas for the student. Practitioners cannot simply assume that any one achievement test is as good as any other. Researchers and test publishers can help psychological examiners by more clearly defining the specific skills related to various achievement constructs so that subtests with the same label (e.g., reading comprehension, numerical operations) more closely assess the same set of skills. Researchers can further assist examiners by clarifying the specific skill deficits that can be expected of students with different disabilities so that examiners can select the achievement test that provides the best measure of these skills.

References

Beck, M. (1992). Review of the KeyMath Revised: A Diagnostic Inventory of Essential Mathematics. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 437-438). Lincoln, NE: University of Nebraska.

Benes, K. M. (1992). Review of the Peabody Individual Achievement Test--Revised. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 437-438). Lincoln, NE: University of Nebraska.

Bracken, B. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26*, 155-166.

Caskey, W. E., Jr. (1985). The use of the Peabody Individual Achievement Test and the Woodcock Reading Mastery Tests in the diagnosis of a learning disability in reading. *Diagnostique, 11*, 14-20.

Caskey, W. E., Jr., Hylton, S., Robinson, J. B., Taylor, R. L., & Washburn, F. F. (1983). Woodcock and PIAT reading scores: A lack of equivalency. *Diagnostique, 9*, 49-54.

Connelly, J. B. (1985). Published tests-which ones do special education teachers perceive as useful? *Journal of Special Education, 19*(2), 149-155.

Connolly, A. (1988). *Key-Math Revised: A Diagnostic Inventory of Essential Mathematics*. Circle Pines, MN: American Guidance Service.

Eaves, R. C., Darch, C., & Haynes, M. (1989). The concurrent validity of the PIAT and WRMT among students with mild learning problems. *Psychology in the Schools, 26*, 261-266.

Finley, C. (1992). Review of the KeyMath Revised: A Diagnostic Inventory of Essential Mathematics. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 438-439). Lincoln, NE: University of Nebraska.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Fuchs, D., & Fuchs, L. S. (1988). Mainstream assistance teams to accommodate difficult-to-teach students in general education. In J. L. Graden, J. E. Zins, and M. J. Curtis (Eds.), *Alternative educational delivery systems: enhancing instructional options for all students*. Washington, D.C.: National Association of School Psychologists.

Markwardt, F. (1989). *Peabody Individual Achievement Test-Revised*. Circle Pines, MN: American Guidance Service.

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-*

based measurement: assessing special children (pp. 18-78). New York: Guilford.

Piotrowski, C., & Keller, J. (1989). Psychological testing in outpatient mental health facilities: A national study. *Professional Psychology: Research and Practice*, 20, 423-425.

Program Standards and Eligibility Criteria for Special Education. (1987). Little Rock, AR: Arkansas Department of Education.

Rogers, B. G. (1992). Review of the Peabody Individual Achievement Test--Revised. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 437-438). Lincoln, NE: University of Nebraska.

Salvia, J., & Ysseldyke, J. E. (1995). *Assessment* (6th ed.) Boston: Houghton Mifflin Company.

Shapiro, E. S., & Derr, T. F. (1987). An examination of overlap between reading curricula and standardized achievement tests. *The Journal of Special Education, 21*(2), 59-67.

Slate, J. R. (1994). WISC-III correlations with the WIAT. *Psychology in the Schools, 31*, 278-285.

Slate, J. R. (1995a). Discrepancies between IQ and Index scores for a clinical sample of students: Useful diagnostic indicators? *Psychology in the Schools, 32*, 103-108.

Slate, J. R. (1995b). Relationship of the WISC-III to the WRAT-R and the PPVT-R for students with academic difficulties. *Assessment in Rehabilitation and Exceptionality, 2*(4), 251-256.

Slate, J. R. (1995c). Two investigations of the validity of the WISC-III. *Psychological Reports, 76*, 299-306.

Slate, J. R. (1996). Interrelationships of frequently administered achievement measures in the determination of Specific Learning Disabilities. *Learning Disabilities Research and Practice, 11*, 86-89.

Slate, J. R., & Saarnio, D. A. (in press). Reading and math achievement test differences for two samples of students with mental retardation. *B.C. Journal of Special Education.*

Webster, R. E., & Braswell, L. A. (1991). Curriculum bias and reading achievement test performance. *Psychology in the Schools, 28*, 193-199.

Wechsler Individual Achievement Test (WIAT). (1992). San Antonio, TX: The Psychological Corporation.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests-Revised.* Circle Pines, MN: American Guidance Service.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 23-34

Preservice Teachers in Two Different Multicultural Field Programs: The Complex Influences of School Context

Janet C. Richards
The University of Southern Mississippi

Ramona C. Moore and Joan P. Gipe
University of New Orleans

This dual site qualitative inquiry offers specific information about two highly regarded university field programs in two different schools that serve culturally diverse children. The study also attempts to make visible the complex influences of each school context on preservice teachers' acquisition of pedagogical content knowledge, concerns and dilemmas, and frames of reference about teaching children in a nonmainstream school setting. Data from the inquiry support the benefits of pluralistic school experiences for future teachers and suggest that contextual conditions unique to a particular school may influence what preservice teachers learn and how they think about teaching.

"The schools in which . . . [preservice teachers are placed] differ in many ways . . . as do the students they teach" (Feiman-Nemser & Floden, 1986, p. 507)

"Anyone who visits schools will be struck by the range of atmosphere or climate they provide" (Sparkes, 1991, p.5)

"When a . . [preservice] teacher enters a school for the first time, she enters more than a building; she enters a culture of teaching" (Bullough, 1987, p. 83)

"To understand the interaction under study, one must also understand the context within which it occurs. This is because . . . the situation can affect perspectives and behavior" (Woods, 1992, p. 358)

Most teacher educators assume the importance of providing pluralistic field experiences for preservice teachers who may have limited views of cultural diversity. Having preservice teachers interact with children whose cultural, linguistic, and home backgrounds differ from their own has the potential to broaden their perspectives, challenge their beliefs, and help them "come face to face with the . . . realities, complexities, difficulties, and rewards of their profession" (Metcalf, Hammer & Kahlich,1995, p. 3). Working in multicultural schools also provides opportunities for preservice teachers to confront, to examine, and, if necessary, to alter their views about teaching diverse or at-risk children (Bondy, Schmitz, & Johnson, 1993). Most importantly, working with children of varying backgrounds may influence preservice teachers from mainstream milieus to develop a willingness for teaching students from diverse cultures and to acquire a commitment and the necessary skills for promoting educational equity in United States' classrooms (Liston & Zeichner, 1990; Ross & Smith, 1992).

Despite current wide acceptance of the benefits of field placements for future teachers, "a substantial and growing body of research suggests that . . . [participating in such programs] may actually lead to less desirable teacher ability" (Metcalf, Hammer & Kahlich, 1995, p. 4).

The concerns of this research fall into four areas: 1) Studies that have examined preservice teachers' experiences in schools generally have relied on indirect measures such as pre- and post-semester surveys and questionnaires rather than "direct, prolonged, on-the-spot [field] observations" (Spindler & Spindler, 1992, p. 63). Few significant indepth studies have been completed.

Janet Richards is an Associate Professor in the Division of Education and Psychology where she supervises field-based literacy methods courses. Ramona Moore recently received a Ph.D. from the University of New Orleans. Joan Gipe holds the rank of Research Professor. The names of the schools and the students' names in this article are pseudonyms. The authors wish to thank the preservice teachers who graciously gave permission for their oral and written comments to be included in the manuscript. An earlier version of this paper was presented at the annual meeting of the American Educational Research Association, New York, April, 1996. Janet C. Richards is Associate Professor, University of Southern Mississippi, Long Beach. Ramona C. Moore is a high school English teacher, Mukelteo School District, Edmonds, Washington. Joan P. Gipe is a Research Professor, University of New Orleans (JPGCI@UNO.EDU). Please address correspondence to Janet C. Richards, The University of Southern Mississippi, 730 East Beach Boulevard, Long Beach, MS 39560-2699 or by e-mail: janetusm@aol.com.

413

2) Practices are based upon unexamined assumptions. Despite acknowledgement of wide diversifications in programs, research that looks at university/K-12 connections generally has excluded descriptions of program characteristics, ignored the specific contexts in which these initiatives take place, and disregarded their impact on preservice teachers' development (Carter, 1990; Feiman-Nemser, 1983; Feiman-Nemser & Buchman, 1987; Guyton & McIntyre, 1990; Liston & Zeichner, 1990; Zeichner, Tabachnick & Densmore, 1987). 3) Participation may have negative effects on preservice teachers' perceptions and practices. For example, because of school culture and norms, characteristics of students, patterns of student/teacher interactions, or how learning is defined, preservice teachers may become preoccupied with group management concerns, may come to consider students with different values, customs, and language as adversaries, and may exhibit less desirable attitudes and performances toward teaching (Evertson, Hawley & Zlotnick, 1985; Richards, Gipe & Moore, 1995); 4) Participation in field placements may influence preservice teachers to develop psychological role conflict and to experience unresolved dissonance. Several studies conclude that placing preservice teachers in classrooms that are incongruous from their own educational and cultural backgrounds negatively affects their self-concepts, motivation for teaching, and sense of self-efficacy (Waxman & Walberg, 1986).

This dual site qualitative inquiry responds to these criticisms by offering context specific information about two highly regarded university field programs in two different schools that serve culturally diverse children. The study also attempts to make visible the complex influences of each school context on preservice teachers' acquisition of pedagogical content knowledge[1], their concerns and dilemmas, and their frames of reference[2] about teaching children in nonmainstream school settings. The inquiry extends previous research in teacher socialization by illuminating preservice teachers' experiences and viewpoints through a predominantly cultural anthropological approach that "builds meaningful generalizations from detailed understandings of specific contexts" (Jacob, 1992, p. 295), and strives to "understand [and interpret] a [social] situation in terms of the individuals involved" (Liston & Zeichner, 1990, p. 611).

Our Purposes for Conducting The Inquiry

As supervisors in charge of field-based literacy methods courses, we spend a considerable amount of time out in the field with our preservice teachers. Because we are close friends and colleagues, we often converse together and share our insiders' knowledge of the elementary schools in which we work. It became apparent to us through our conversations that the contextual conditions of the two schools differed considerably. Yet, we had no specific documentation to support these conclusions. We also had strong hunches that different contextual variables associated with each school influenced our preservice teachers' professional thinking and development in both positive and negative ways, but again we lacked definitive information to substantiate our hunches.

Our purposes for conducting the inquiry were to illuminate the complex social situations and interactions within each school and to attempt to discover the preservice teachers' conceptualizations of their school experiences (Schempp, Sparkes & Templin, 1993). Once we grasped a better understanding of the contextual realities of the two schools and "the demands and expectations pressed upon . . .[our preservice teachers] by [each] school" (Veeneman, 1984 in Schempp, Sparkes & Templin, 1993, p. 448), we could modify course content, if necessary, and offer special activities to help our preservice teachers recognize the linkages between their teaching situations and their own perspectives and educational practices (Liston & Zeichner, 1990). Thus, through our research efforts, we hoped to provide maximum learning conditions for our preservice teachers that might shape in positive ways, their future thinking and actions as classroom teachers.

The Elementary School Contexts

Diamond Elementary School

The first university program meets in a K-8 urban school in New Orleans, Louisiana. Diamond Elementary School is located in a very old, three-story, non-air-conditioned, red brick building. Paint peels from the walls, light bulbs hang suspended from frayed cords, classroom ceilings leak during rain storms, and the hallways and stairwells are dark and dingy. In the spring and fall, temperatures in individual classrooms may reach over 100 degrees. Apparently the school board does not consider Diamond when allocating money for city-wide school improvements.

---

[1] Pedagogical content knowledge is conceptualized as teachers' knowledge of a specific content, and includes beliefs about why and how to teach, and knowledge of the learner (Shulman, 1986; Stein, Baxter & Leinhardt, 1990).

[2] Teachers' frames of reference are defined as unique to the individual, born of experience, and indicative of a teacher's values, thoughts, "ideas and perspectives at a single point in time" (Mager, Alioto, Warchol & Carapellas, 1995, p.9). Frames of reference are very practical, and include the knowledge, values, and skills that will shape each preservice teacher's future thinking, and what they will be inclined to do as classroom teachers (Mager, Alioto, Warchol & Carapellas, 1995).

There is a permissive atmosphere in this school and an optimism concerning students' abilities and motivations to be responsible, motivated, and self-directed. The 300 students (20% Caucasian; 80% African-American), address teachers by their first names, and they are allowed to walk out of classrooms without asking permission in order to use the bathroom, or to go to the water fountain, or to speak with the principal concerning problems with peers or teachers. They also are encouraged to interact and verbalize with one another whenever they wish. Consequently, the noise level is high. Many students in grades two and above are over age for their grade placement, and many of the older students were dropouts for a year or more prior to attending Diamond. "Unable to achieve in school, these . . . [students] . . . see academic success as unattainable and so they protect themselves by deciding school is unimportant" (Comer, 1988, p. 6). Unfortunately, a sense of inadequacy, low self-esteem, inner conflicts, chronic anger, and peer pressures contribute to some students exhibiting developmentally inappropriate, disruptive behavior, such as fighting, walking out of classrooms, running through the halls, talking out during lessons, verbally challenging their teachers, and occasionally deliberately trying to offend the university preservice teachers (e.g., "I hate white people!"). Since student individuality and freedom of expression are stressed, few students receive reprimands or consequences for inappropriate actions or comments.

Each semester a few of the older girls become pregnant, and a few of the older boys are expelled for illegal activities such as selling or using drugs or carrying concealed weapons. In one recent, isolated incident, an eighth grade boy superficially wounded himself in the knee when he reached into his bookbag and accidentally discharged a pistol that he had brought to school to use as a defense against a gang of bullies in his neighborhood.

Pedagogical autonomy is offered to the 13 teachers (eight Caucasian and five African-American), at Diamond Elementary School, and their instructional orientations range from a "teacher-as-information-giver" view to a constructivist, "teacher-as-facilitator" approach. Most of the students' reading and language arts test scores are significantly below the norm.

Forest Park Elementary School

The second university program is located in a modern, beige brick, single-story, air-conditioned K-6 school in a small town on the Mississippi Gulf Coast. Forest Park Elementary School is exceptionally clean, orderly, and quiet. An authoritarian, inflexible, custodial attitude permeates Forest Park. The 20 teachers (6 Caucasian; 14 African-American), are definitely in charge. Control of the 450 students (5% caucasian; 95% African-American), and regimentation are valued and stressed. Students follow their teachers' directions without question. For the most part, they are expected to work silently and alone. Classes walk in silent lines through the halls when they are going to physical education or music lessons, and some teachers tell students to hold their fingers over their lips as a reminder that talking is forbidden. Students must address teachers by their formal names, and they are expected to answer teachers' questions with a nod of their head followed by a "Yes, ma'am" or "No, ma'am." Students are punished for breaking even the smallest rule by being sent to a special in-school suspension room, where they work silently under the direction of a full time teachers' assistant. Students call this type of punishment "being in the hot seat." Paralleling students' affective dimensions at Diamond, many students at Forest Park also experience feelings of inadequacy, low self-esteem, inner conflicts, chronic anger, and pressure from peers. Because of fear of punishment, they seldom manifest their feelings by "acting out." When they do exhibit inappropriate behavior, they are immediately isolated.

The instructional philosophy of all 20 teachers at Forest Park is "teacher-as-information-giver." Effective teaching is equated with keeping students quiet and on-task. Lessons are decontextualized and skills-based, and students work on only one assignment at a time (e.g., reading from 9:00 to 9:50 AM). Here, as in Diamond, the majority of the students' reading and language arts standardized test scores are significantly below the norm.

Commonalities of Both University Programs

Four commonalties undergird both university programs: 1) Guided by a constructivist, inquiry-oriented view of learning, both university programs focus on integrated, literature-based instruction; 2) As teacher educators in charge of both programs, we hold holistic, student-centered orientations; 3) The university course activities of both programs are similar; and 4) We work to insure that we offer the preservice teachers in both programs ongoing, detailed supervision and caring, empathic support.

Research Methodology

Theoretical Perspective

Tenets of qualitative inquiry guided our research. Qualitative methods are especially appropriate when research is field-focused (Eisner, 1991) and when researchers wish to provide "rich descriptive data about the contexts, activities, and beliefs of participants in educational settings" (Goetz & LeCompte, 1984, p. 17). Four literatures informed the study: 1) Perspectives from symbolic interactionism that consider the social organizations of schools and help to explain why school organizational structures influence students' and teachers' actions (Alvermann, O'Brien & Dillon, 1996; Erickson, 1992; Measor & Woods, 1984; Woods, 1992); 2) Research which suggests that norms, traditions, roles, and values are crucial contextual variables, and that beliefs, behavior, and learning are "socially constructed in the course of interaction with students, teachers, and others" (LeCompte & Preissle, 1992, p. 818); 3) Ideas from sociology, anthropology, and sociolinguistics that emphasize human development in terms of culture and context (Grant & Fine, 1992; Spindler & Spindler, 1992); and 4) Traditions from hermeneutic interpretive analysis that provide "a window for viewing [preservice] teachers' renditions of their in-school experiences" and which can reveal and identify "the contextual social rules that underlie [preservice] teachers' actions" (Schempp, Sparkes, & Templin, 1993, p. 449).

We designed the study "following guidelines of participant-observational field work [with the goal of gaining insights into] the meaning-perspectives of [the preservice teachers]" (Rovegno, 1992, p. 71). In addition, we were sensitive to the construct of triangulation since "the act of bringing more than one source of data to bear on a single point . . . [and] designing a study in which multiple cases, multiple informants or more than one data gathering are used can greatly strengthen the study's usefulness for other settings" (Marshall & Rossman, 1989, p. 146).

Study Participants

Study participants were 85 female and three male junior or senior preservice teachers, majoring in elementary or exceptional education (45 preservice teachers were at Diamond Elementary School, and 43 preservice teachers were at Forest Park Elementary School). Of these 88 preservice teachers, 85 were Caucasian, two were African-American, and one was Hispanic-American. All were from middle socioeconomic backgrounds. Their ages ranged from 21 to 42 years. The preservice teachers had no prior teaching experience and were enrolled in fall or spring semester, required, reading and language arts, field-based courses offered in one of two colleges of education located in adjacent southern states.

Data Sources and Data Analysis

Serving as participant researchers (Taylor & Bogdan, 1984), we collected data for two semesters in both school settings. Data sources were field notes of teaching observations and conversations as well as artifacts--"texts which themselves are implicated in the everyday construction of reality" (Atkinson, 1990, p. 178). The artifacts included preservice teachers' dialogue journals, metaphors and semantic maps depicting their teaching experiences, and final reflective statements (see Appendices A through D for examples of these artifacts). Additionally, we photographed ongoing activities of both programs in order to capture in an objective way the events particular to each school setting and the "daily life of the group[s] under study" (Marshall & Rossman, 1989, p. 86).

At the end of each semester we met together as a research team, collating all of the data sets for study participants, making notes, and transcribing the data when necessary (e.g., transcribing our understandings of the teaching/learning activities and the social interactions depicted in the photographs). We also made two separate listings of the items mentioned in the preservice teachers' semantic maps according to whether they worked at Diamond or at Forest Park Elementary School, and then we tallied and compared the frequency of the items on each list (see Appendix E for a listing of the six most frequently mentioned items on the preservice teachers' semantic maps).

In subsequent meetings, we conducted content analyses, comparing and cross-checking the aggregated data in order to identify and code similar themes and patterns (Glaser & Strauss, 1967). We negotiated points of disagreement through roundtable discussions until we reached a consensus.

Methodological Limitations

Methodological limitations to the inquiry must be acknowledged. First, as is common to all research efforts, we brought our own backgrounds to the inquiry and our own particular ways of interpreting the data. "Qualitative researchers can never overlook the fact that they are gendered, multiculturally situated, and theoretically inclined to view phenomena in ways that influence what questions get asked and what methodology is used to answer those questions" (Denzin, 1994 in Alvermann, O'Brien & Dillon, 1996, p. 116). "There is always the question of [subjective] interpretation [in social research]" (Blachowitz & Wimett, 1994, p. 11). However, the credibility or "truth value" of our efforts was established through structural corroboration (i.e., "the use of multiple sources and types of data to support [our] . . . interpretations") (Pitman & Maxwell, 1992, p. 748). Second, our interpretive approach relied on the subjective understandings and points of view of 88 preservice teachers. It is highly possible that other preservice teachers might hold different perceptions and understandings about their teaching experiences in the two schools. Third, the inquiry examined two public elementary schools where "teaching takes place in a complex social situation, with each . . . school . . . offering a distinct constellation of social conditions. Generalizable principles . . . are simply not possible" (Schempp, Sparkes & Templin, 1993).

Major Themes Emerging from the Inquiry

Twelve major themes emerged from the inquiry which we accept as representing the preservice teachers' realities at a single point in their professional careers. The themes illuminate the distinct characteristics of the two elementary schools and suggest that certain contextual variables unique to each field placement influenced the preservice teachers' initial impressions about their teaching assignments, feelings of frustration or confidence, group management concerns, regard for children's well-being and learning, acquisition of pedagogical content knowledge, and sense of success or failure.

Examination of the Themes

1. The two schools differ considerably with respect to classroom teachers' expectations for students' behavior, student/teacher interactions, and how learning and school achievement are defined. These expectations, norms, and traditions were conveyed through both formal and indirect means to the preservice teachers, and influenced their initial impressions about the schools in which they were placed (e.g., ***Diamond Elementary School***: "Nothing could have prepared me for what I saw today. All these kids do is talk and the teacher just lets them do whatever they want." * "I am really worried. I heard about how the kids run this school and now I know it's true"; ***Forest Park Elementary School***: "I made a cake to use as a visual with our story and when I served it to the kids the teacher ran right over because she doesn't allow them to do anything like eating cake or having fun in school."* " I feel like I'm in a strange country and I don't know the rules but I'm sure I'll break them." * "I know I'll break a rule and I feel like the rules keep changing and expanding." * "What are the rules today?").

2. Preservice teachers in both colleges of education entered the field placements sharing unrealistic expectations regarding their abilities to influence students' learning and holding idealized conceptions about teaching (e.g., ***Diamond Elementary School***: "I will work to make sure that all of my students learn how to read and write"; ***Forest Park Elementary School***: "I know I'll be able to help them become turned on to school and to learning").

3. All 88 preservice teachers experienced initial anxieties and feelings of vulnerability about teaching in a nonmainstream school setting (e.g., ***Diamond Elementary School***: "I have fears. I was awake all night worrying about going to that school."* "I had heard horror stories and they are true."; "I have memorized the route to and from the school so that I won't get off the main roads."; ***Forest Park Elementary School***: "Why were we placed in this school?" * "I am uptight and anxious. I can't even pronounce these children's names.").

4. Twelve preservice teachers in the fall semester and 10 preservice teachers in the spring semester at Diamond Elementary School developed negative frames of reference about teaching children in a culturally diverse school setting (e.g., "It was hard to relate to these students since I do not come from where they do. I'll apply for a position in my own community first."* "Why don't these children act like the children where I teach preschool? When I tell my preschool children to line up, they line up!" * "This place is shocking!"). Three preservice teachers in the fall semester and two preservice teachers in the spring semester at Diamond Elementary School developed positive frames of reference about teaching students in a culturally diverse school setting (e.g., "These students taught me a lot about inner city children who I knew nothing about."). In contrast, 15 preservice teachers in the fall semester and 22 preservice teachers in the spring semester at Forest Park Elementary School developed positive frames of reference about working with culturally diverse students (e.g., " I love these children and I just wish that I could help all of them.")

5. Similar to Grossman and Richert's (1986) conclusions regarding the connection between field experiences and preservice teachers' knowledge development, participation in either school placement appeared to facilitate the preservice teachers' acquisition of pedagogical content knowledge in positive ways. However, the preservice teachers at Diamond Elementary School focused more on their own learning while the preservice teachers at Forest Park Elementary School focused more on what individual children were learning (e.g., ***Diamond Elementary School***: "I would say my greatest

success was in learning how to teach reading and language arts."* "I have begun to solidify my views about teaching reading."; ***Forest Park Elementary School***: "I initially chose books above their level."*"He is doing much better. He wrote two whole sentences in his journal today.").

6. Seven preservice teachers in the fall semester and 12 preservice teachers in the spring semester at Diamond Elementary School continued to experience anxieties and frustrations throughout the semester. They manifested their anxieties by distancing themselves from their students and by complaining (e.g., "There are only three more weeks in this semester and I am out of here. Thank God!"* "Regretfully, I must tell you that I hated this place."*"I have been in this school a month now and I am losing the war."*"The student that I enjoy working with the most is the only Caucasian student in my group. I can communicate with him."*"I have finally wised up! I just send him out of the group. When I send him away he just says, 'So!'").

7. Contrary to the findings of other studies which suggest that for the most part, field experiences promote preservice teachers' authoritarian perspectives, (e.g., see Zeichner, 1980), all 43 of the preservice teachers at Forest Park Elementary School developed student-centered orientations (e.g., "Those poor kids--they can't even talk. I'll never be a teacher who is so strict that kids can't enjoy school.").

8. Three preservice teachers in the fall semester and two preservice teachers in the spring semester at Diamond Elementary School experienced ongoing feelings of extreme failure (e.g., "I feel like I have failed my profession."* "I am so sorry and regretful that I never really did a good job.").

9. Similar to conclusions reported by Hoy and Woolfolk (1990) in their study of the effects of organizational socialization on student teachers' preoccupations with student management, the majority of preservice teachers at Diamond Elementary School remained preoccupied throughout the semester with group management concerns. (e.g., "Next week I'm not going to let them say one word out of turn!"* "I am going to try a token system so that I can reward students for good behavior, *if there is any good behavior*.").

10. The preservice teachers at Diamond Elementary School mainly were concerned about classroom teachers' permissive attitudes and "unruly," unmotivated students (e.g., "All of the problems were made worse by the classroom teacher's attitude about not letting me give kids time-out."* "These students are constantly disruptive, hostile, boisterous, and disinterested. They also waste time.") The preservice teachers at Forest Park Elementary School mainly were concerned about offering effective literacy lessons, enhancing individual student's social, emotional, and academic well-being, and circumnavigating classroom teachers' controlling orientations (e.g., "After pondering and worrying about the situation for an entire week, I decided to continue to take her dictation. She still is insecure about writing."* "I am going to get a book about knights and dragons since he has no knowledge of these and he confuses the word 'knight' with 'night'." * "I am concerned because of his home life." * "I felt like our drama practice was being monitored by the teacher. I am determined to practice with my kids in privacy.").

11. Over the course of the semester, the majority of preservice teachers in both schools came to value their teaching experiences and developed confidence in their teaching abilities. (e.g., ***Diamond Elementary School***: "I now know more than any outsider would ever believe. I can teach anywhere."* "I'm going to be a great teacher!"; ***Forest Park Elementary School***: "Every future teacher needs this experience. I feel like I'm more than ready for student teaching.").

12. There are strengths and shortcomings associated with both field contexts (e.g., ***Diamond Elementary School***: "One of the problems in this school is their overly student-centered philosophy."; ***Forest Park Elementary School***: "One of the problems in this school is the philosophy of strict-no talking.").

Our Conclusions

As thoughtful, qualitative researchers, we know that caution must temper our conclusions. Contextual considerations are "subtle and difficult to [capture and] untangle" (Hoy & Woolfolk, 1990 p. 296). Context is "experienced, understood, and interpreted [individually]" (Clarke, Hall, Jefferson, & Roberts, 1981, pp. 52-53) and therefore, is one of the most complex, and elusive qualities to understand and describe (Gibson, 1986). Nonetheless, we conclude that the inquiry provides several important considerations. First, working with minority students in either school setting appeared to enhance considerably the preservice teachers' professional thinking and development (e.g., "I sure learned about myself as a professional. I did a lot of growing up this semester, and I realize I still have a long way to go.")

Second, despite some of the preservice teachers' concerns, anxieties, and ongoing feelings of failure, the majority of preservice teachers in either field placement came to recognize the value of their school experiences, and they developed confidence in their abilities to teach students from diverse backgrounds (e.g., "This teaching experience is my first of actually working with children on a continuous basis. In the first few weeks I was terrified. I attended private schools and my first glimpse of an urban school shocked me. Only two of my students in second grade can read. But, a lot of people ask me how I like teaching in 'that part of town' and you know what? -- I can't shut up about MY second graders.").

The inquiry also suggests that contextual conditions unique to a particular school may play a part in influencing what preservice teachers learn and how they think about teaching. For example, the preservice teachers at Diamond Elementary School (the permissive, student-centered school) developed authoritarian attitudes toward their students (e.g.,"I will continue to go over the rules and send kids away from the group."). In contrast, the preservice teachers at Forest Park Elementary School (the authoritarian school) shifted toward student-centered orientations (e.g., "Is there anything these kids are allowed to do? They can't even paint or sing."). Another difference is that the foremost concerns of the preservice teachers at Diamond Elementary School centered around classroom teachers' permissive attitudes and "unruly" students (e.g., "She never reprimands them!!" * "I am thrilled when William is absent. He is the worst to manage."), while the concerns of the preservice teachers at Forest Park Elementary School focused more on their effectiveness as literacy teachers, individual student's learning and well-being, and classroom teachers' authoritarian attitudes (e.g., "She told me that she has no father and her mother isn't home very much. Isn't that terrible? -- the poor child." * " Now they won't even let us take our kids out in the hallway for a minute."). Another intriguing finding is that while the preservice teachers in both colleges of education recognized how much they had learned about teaching reading and language arts, the preservice teachers at Diamond Elementary School focused more on their own learning (e.g., "I never knew about literature-based instruction."), while the preservice teachers at Forest Park Elementary School focused more on individual students' literacy progress (e.g., "He wrote back to me in his journal for the first time. I'm thrilled.").

We conclude that these phenomena occurred because of the different realities associated with each school. It appears that concerns with managing groups of students influenced the preservice teachers at Diamond Elementary School to become more controlling in their orientations toward students as a means of accomplishing their teaching goals and successfully completing course requirements. Preoccupation with maintaining group order also prohibited these preservice teachers from interacting with individual students, and getting to know individual students' strengths and specific academic and emotional needs. Understandably, these preservice teachers concentrated more on their own learning and development as a way to assuage their struggles and affirm their worth as beginning professionals. In contrast, the preservice teachers at Forest Park Elementary School had very few problems with student behavior. Therefore, they had ample opportunities to get to know students on a personal level and considerable time to reflect about individual students' academic and emotional needs. In addition, it is quite clear that these preservice teachers resented and consciously resisted Forest Park's custodial atmosphere. We surmise that they adopted student-centered perspectives as a way to counteract the classroom teachers' strict attitudes.

Conducting the inquiry has provided information that supports our convictions about the benefits of diverse school experiences for future teachers. At the same time, engaging in the research has pushed us to reevaluate our field programs. We now have substantial evidence that the two schools differ, and we are aware of some of the contextual strengths and shortcomings of each school. Therefore, we understand the importance of spending time selecting field sites with characteristics that are congruent with the goals and orientations of our preservice teacher preparation curriculums (Zeichner & Liston, 1987). In addition, we know that school experiences can exert both positive and negative consequences on preservice teachers' professional thinking and development. Consequently, we recognize that it is imperative for us to monitor our preservice teachers' perceptions, concerns, and professional growth throughout their time in the field so that we can offer appropriate interventions whenever necessary. Finally, we conclude that we need to refashion our programs. We need to provide opportunities for our preservice teachers to get to know their students as individuals with different interests, perspectives, and strengths. We need to include activities such as case writing and action research projects that can help our preservice teachers learn how to look purposefully and critically at their teaching and how to brainstorm and problem-solve effective ways for reaching their students. We need to provide seminar discussions and readings about multi-cultural issues that can help our preservice teachers analyze the contextual conditions of a school and come to understand how these conditions have the capacity to shape their teaching views, their pedagogical actions, and ultimately, their teaching successes (Sparkes, 1991). These program changes could very well insure that all of our preservice teachers develop positive attitudes toward their experiences in the field, as well as the knowledge and commitment necessary for effectively teaching culturally diverse children.

References

Alvermann, D., O'Brien, D., & Dillon, D. (1996, January, February, March). On writing qualitative research. *Reading Research Quarterly, 31*(1), 114-120.

Atkinson, P. (1990). *The ethnographic imagination: Textual constructions of reality*. New York: Routledge.

Blachowitz, C., & Wimett, C. (1994). *Reconstructing our pasts: Urban preservice teachers' definition of literacy and literacy instruction*. Paper presented at the Annual meeting of the National Reading Conference, New Orleans, LA.

Bondy, E., Schmitz, S., & Johnson, M. (1993). The impact of coursework and fieldwork on student teachers' reported beliefs about teaching poor and minority students. *Action in Teacher Education, 15*(2), 55-62.

Bullough, R. (1987). Accommodation and tension: Teachers, teacher role, and the culture of teaching. In J. Smith (Ed.), *Educating teachers: Changing the nature of pedagogical knowledge* (pp. 83-94). Lewes, England: Falmer Press.

Carter, K. (1990). Teachers' knowledge and learning to teach. In W. Houston (Ed.), *Handbook of research on teacher education* (pp.291-310). New York: Macmillan.

Clarke, J., Hall, S., Jefferson, T., & Roberts, B. (1981). Subcultures, cultures and class. In T. Bennett, G. Martin, C. Mercer, and J. Woollacott (Eds.), *Culture, ideology and social process* (pp. 53-79) London: Batsford.

Comer, J. (1988, November). Educating poor minority children. *Scientific American, 259*, 2-8.

Denzin, N. (1994). Evaluating qualitative research in the poststructural moment. The lessons James Joyce teaches us. *Qualitative Studies in Education, 7*, 295-308.

Eisner, E. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Erickson, F. (1992). Ethnographic microanalysis of interaction. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp.201-225). New York: Academic Press.

Evertson,C., Hawley, W., & Zlotnick, M. (1985). Making a difference in educational quality through teacher education. *Journal of Teacher Education, 36*(3), 2-12.

Feiman-Nemser, S. (1983). Learning to teach. In L. Shulman, and G. Sykes (Eds.), *Handbook on teaching and policy* (pp. 150-170). New York: Longman.

Feiman-Nemser S., & Buchman, M. (1987). When is student teaching teacher education? *Teaching and Teacher Education, 3*, 255-273.

Feiman-Nemser, S., & Floden, R. (1986). The cultures of teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp.505-526). New York: Macmillan.

Gibson, R. (1986). *Critical theory and education*. London: Hodder and Stoughton:

Glaser, B., & Strauss A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. NewYork: Aldine De Gruyter.

Goetz, J., & LeCompte, M. (1984). *Ethnography and qualitative design in educational research*. Orlando, FL: Academic Press.

Grant, L., & Fine, G. (1992). Sociology unleashed: Creative directions in classical ethnography. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 405-446). New York: Academic Press.

Grossman, P., & Richert, A. (1986). Unacknowledged knowledge growth: A re-examination of the effect of teacher education. *Teaching and Teacher Education, 41*(1), 53-62.

Guyton, E., & McIntyre, D. (1990). Student teaching and school experiences. In W. Houston (Ed.), *Handbook of research on teacher education* (pp. 514-534). New York: Macmillan.

Hoy, W., & Woolfolk, A. (1990, Summer). Socialization of student teachers. *American Educational Research Journal, 27*(2), 279-300.

Jacob, E. (1992). Culture, context, and cognition. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 293-335), New York: Academic Press.

LeCompte, M., & Preissle, J. (1992). Toward an ethnology of student life in school and in classrooms: Synthesizing the qualitative research tradition. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education*. (pp. 816-859). New York: Academic Press.

Liston, D., & Zeichner, K. (1990, Winter). Teacher education and the social context of schooling: Issues for curriculum development. *American Educational Research Journal, 27*(4), 610-636.

Mager, G., Alioto, P., Warchol, M., & Carapellas, F. (1995, Winter/Spring). Upon which to build: Developing a base for practice as a teacher educator. *Teaching Education, 7*(1), 7-13.

Marshall, C., & Rossman, G. (1989). *Designing qualitative research.* Newbury Park, CA: Sage Publications.

Measor, L., & Woods, P. (1984). *Changing schools: Pupils perspectives on transfer to a comprehensive.* Milton Keynes, England: Open University Press.

Metcalf, K., Hammer, R., & Kahlich, P. (1995, April). *Alternatives to field-based experiences: The comparative effects of on-campus laboratories.* Paper presented at the Annual meeting of the American Educational Research Association, San Francisco, CA.

Pitman, M., & Maxwell, J. (1992). Qualitative approaches to evaluation: Models and methods. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 729-770). New York: Academic Press.

Richards, J., Gipe, J., & Moore, R. (1995). Lessons in the field: Context and the professional development of university participants in an urban field placement. *Research in the Schools*, 41-54.

Ross, D., & Smith, W. (1992). Understanding preservice teachers' perspectives on diversity. *The Journal of Teacher Education, 43*(2), 94-103.

Rovegno, I. (1992). Learning to teach in a field-based methods course: The development of pedagogical content knowledge. *Teaching and Teacher Education, 8*(1), 69-82.

Schempp, P., Sparkes, A., & Templin, T. (1993, Fall). The micropolitics of teacher induction. *American Educational Research Journal, 30*(3),447-472.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Sparkes, A. (1991). The culture of teaching, critical reflection and change: Possibilities and problems. *Educational Management and Administration, 19*(1), 4-19.

Spindler, G., & Spindler, L. (1992). Cultural process and ethnography: An anthropological perspective. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 54-92). New York: Academic Press.

Stein, M., Baxter, J., & Leinhardt, G. (1990, Winter). Subject matter knowledge and elementary instruction: A case from functions and graphing. *American Educational Research Journal, 27*(4), 639-663.

Taylor, S., & Bogdan, R. (1984). *Introduction to qualitative research methods: The search for meaning.* New York: John Wiley.

Veeneman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*(2), 143-178.

Waxman, H., & Walberg, H. (1986). Effects of early field experiences. In J. Raths and L. Katz (Eds), *Advances in teacher education*, Vol 2 (pp. 165-184). Norwood, NJ: Ablex.

Woods, P. (1992). Symbolic interactionism: Theory and method. In M. LeCompte, W. Millroy, and J. Preissle (Eds.), *The handbook of qualitative research in education* (pp.337-404), New York: Academic Press.

Zeichner, K. (1980, November-December). Myths and realities: Field-based experiences in preservice teacher education. *Journal of Teacher Education, 31*(6), 45-55.

Zeichner, K., & Liston,. D. (1987). Teaching student teachers to reflect. *Harvard Educational Review, 57*, 23-48.

Zeichner, K., Tabachnick, B., & Densmore, K. (1987). Individual, institutional, and cultural influences on the development of teachers' craft knowledge. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 21-59). London: Cassell Educational Ltd.

Appendix A

Examples of Preservice Teachers' Metaphors about Teaching

Diamond Elementary School

- "Teaching here is like being in a three-ring circus. There is chaos everywhere. In each ring there is a new and somewhat bizarre escapade, only I'm not laughing."

- "My experiences here have been like holding a rose. Like a rose, it's appearance seems wonderful, but the minute you hold that rose wrong the thorns will cut your hand. On some days my children are the sweet beautiful petals but on other days they are the dreadful thorns that cut into my side."

- "Teaching at this school is like being thrown to the wolves. Sometimes I feel as though I was just thrown into teaching and that I have no idea of what I am doing. I am unsure if my students are learning anything."

- "Teaching at this school is like walking through a maze. Before you go into the maze you feel somewhat confident but once you're in the maze you feel disoriented and confused."

Forest Park Elementary School

• "Teaching here this semester is like being a foreigner in a strange land where you don't know the rules but you are sure just as soon as you do something wrong, someone will correct you."

• "Teaching here is like digging for buried treasure. I know something is there but it's been covered for so long you can't get to it."

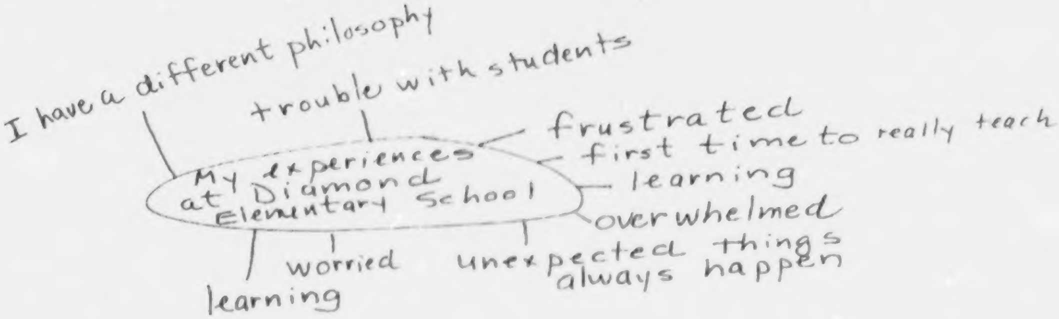
• "Teaching here is like wanting to fly and realizing your wings have been clipped."

• "Teaching here is like being afraid to take a chance because the headmaster might crack his whip."

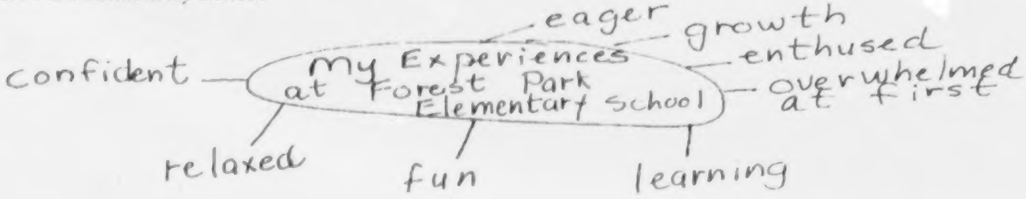
Appendix B

Examples of Preservice Teachers' Semantic Maps Depicting Their Teaching Experiences

Diamond Elementary School


My experiences at Diamond Elementary School
I have a different philosophy
trouble with students
frustrated
first time to really teach
learning
overwhelmed
unexpected things always happen
worried
learning

Forest Park Elementary School


My Experiences at Forest Park Elementary School
eager
growth
enthused
overwhelmed at first
confident
relaxed
fun
learning

BEST COPY AVAILABLE

Appendix C

Examples of Preservice Teachers' Final Reflective Statements

Diamond Elementary School

Wow! What a semester. I hardly know where to begin as I look back and reflect on the work I did. I've had good days and bad days. At the beginning of the semester, I felt overwhelmed. I felt as if I had been thrown into a teaching position for which I was not ready. I had no idea how to teach, or how to help students learn about reading and language arts. I was somewhat familiar with the phonics approach--that's it! I remember feeling at a loss, wishing that I knew some strategies or methods to help the students learn. I do not know when I thought I would learn these strategies unless it was a year-long class (which isn't a bad idea!).

During the semester there were times when my work with the kids was extremely stressful and at other times it was notably joyous. The children were not at all what I expected. They all contributed to the lessons at various times. But, they also had their moments when they disturbed others and refused to participate. I tried to have good communication with my kids. I believe that this enabled me to get to know them better.

One of the struggles I had to deal with was a student named Ben. I just could not get his attention. I do not know if I was doing something wrong or if he was the one with the problems. But, from working with him I learned a few strategies that I can use in similar situations in the future.

I sure learned a lot about myself as a professional. I did a lot of growing up this semester and I realize I still have a long way to go. I've learned about group management techniques---some worked well and some didn't. I am glad that I had an opportunity to explore and experiment with these different techniques and find out which ones work for me. I've learned that I have to be consistent and fair to all of my students and that I must follow through with set consequences. I also need to be over prepared when I walk into a classroom. Another area I think I've improved in is that I learned how to relax a little bit with students.

You probably want to know about our failures. Well, it sure was a failure with Ben. I was never able to help him. In addition, many of the reading strategies you taught us were above these students' heads--that includes prediction logs, literature logs, cloze passages, and basic story features and their connections. A big surprise to me was the mural. Despite everything that I had heard from other former preservice teachers, I was sure that it would be a disaster. But, I was wrong. The kids loved it, cooperated, and were angels. I should have done more murals.

I certainly will take with me the idea that all children have a right to the best possible education they can get, despite their family income. There is also something that I will not take with me and that's the "free to be me" approach in this school.

Finally, I can honestly say that I am thankful for this experience. I do not picture myself ever teaching in this school. There were days when I was really stressed out. But I did learn a lot about myself as a professional, my teaching philosophy, the type of group management system I like, and ideas for having a student-centered classroom.

This class was very overwhelming for me at the beginning. I didn't know what was going on and I thought that I'd never be able to do it all. As time went on, and after asking tons of questions, I started to understand. I think what helped me the most was the demonstration lessons you presented. I never realized how much there was to teaching--like before you even read the story. All in all, I'm thankful for the experience.

Forest Park Elementary School

As I look back on the semester I can't even count the number of different feelings and emotions that I have felt. I have gone from fear to confusion to enthusiasm to joy. I have learned so much from this class. The children taught me so much. I now know what to expect from them. Not only are they capable of doing this work, they do a wonderful job. They need plenty of attention and encouragement while doing their work, but, isn't that my job as a teacher? This experience has taught me to expect the best from my students and to love them. All teachers should love their students and treat them kindly. Some of the teachers here should learn to love their kids. If they really talked to their kids, they would see that every one of them is special.

As I began this semester, I felt like I could save the world and I wanted to save the world. I thought, "I will dedicate my time to Jimmy and other students like him." Now I realize that I will not be able to do this. I can still help but I cannot perform miracles.

I have recently come to realize that attention is the power of the teacher. I have it to give and the students need it. For example, Andeus begins acting up when my back is turned, however, the minute I start working with him, he becomes an angel. These students get so little attention at home and in this school that they crave individual attention.

I guess the best thing I have learned from you and this experience is that praising children is the best medicine for any student. The response is overwhelming. Children will do so much for so little. Praise is an excellent tool and should be mastered by every teacher. You are an outstanding role model for this technique and I hope that soon all of the teachers in this school will recognize the wonderful things that happen when the teacher is student-centered and when teachers give students love, attention, and praise. For this experience I say, "Thank you."

Appendix D

Examples of Preservice Teachers' Dialogue Journal Entries

Diamond Elementary School

Dr. R.---I am stressed out. These kids don't listen and they don't like my lessons. I think they don't like me either. What suggestions do you have to help me get these students to listen? When I get ready to go into the classroom I can hear them all talking and making noise. Even the teacher can't seem to do anything with them. Why do we teach here?

Forest Park Elementary School

This school is very strict on kids. How will they learn to give their opinions if they can't talk and share ideas? Also, drama practice is really getting me upset. The teacher monitors our drama practice and so I feel that we get nothing accomplished because we're uptight. These rules are killing me. How can we follow the rules when they keep on adding new ones?

Appendix E

A Listing of the Six Most Frequently Stated Items on the Preservice Teachers' Semantic Maps About Their Teaching Experiences

Diamond Elementary School	*Forest Park Elementary*
1. overwhelmed	1. learning
2. frustrated	2. fun
3. excited	3. relaxed
4. learning	4. confident
5. scared	5. anxious/challenged/ enjoyable/ overwhelmed
6. problems with group management	6. valuable/enthused

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 35-39

Beam Me Up Scottie: Professors' and Students' Experiences With Distance Learning

Neelam Kher-Durlabhji and Lorna J. Lacina-Gifford
Northwestern State University

Distance learning is becoming an integral part of higher education because of increasing demands for advanced training. This study focuses on distance learning instruction in graduate level courses from the perspectives of three faculty members and their students in the studio and at remote sites. Information is provided on the nature and description of courses being taught, preparation of faculty for teaching by satellite, impressions of being "live" on camera, modifications to the instructional process because of alternate delivery, and an evaluation of the satellite teaching experience. Differences between the perceptions of students in the studio and at remote sites is also discussed. This study identifies a number of issues that will require research and debate as distance learning becomes more prevalent in higher education.

Distance learning is fast becoming a growing alternative for traditional education. It is defined as "the use of telecommunications equipment such as the telephone, television, fiber optics, cable broadcast, and satellites to send instructional programming to learners" (Bruder, 1991, p. 20). It includes anything from correspondence courses to live interactive instruction.

Satellite-based programs depend on one-way video and two-way audio for student interaction. The only site that is actually seen by all students is the studio location, but the audio connection is two way, which enables the instructor to answer questions on the air for all students to hear (Ostendorf, 1989). While studies and commentaries on distance learning appear to be voluminous, in actuality there are large and disappointing gaps, especially in higher education (Speth, 1990).

Distance learning also includes "network-focused distance learning" which utilizes microcomputers, modems, electronic bulletin boards and e-mail (Barker & Hall, 1993); computer mediated communication which uses personal computers, modems, phone lines, and computer networks as tools for group communication and cooperative learning (Schrum, 1992); computer conferencing which is live instruction accessed via a microcomputer and modem (Eastman, 1994); and electronic classrooms where students check into their seats via modem, receiving batch information in their e-mail accounts, read a lecture, participate in discussions, and take tests (Mizell & Carl, 1994). The latest type of distance learning is learning by means of the internet. In an internet course, students receive information via e-mail and electronic bulletin board, participate in discussions via a bulletin board on issues, conduct research through the net, and e-mail assignments to the instructor (Lacina-Gifford & Kher-Durlabhji, 1995).

More and more distance learning students are working adults who want a specialization or a degree (Hyatt, 1992). The greatest benefits of distance learning indicated by students include convenience, time flexibility, and decreases in travel time (Hyatt, 1992).

According to Lever (1993), distance learning will inevitably become an integral part of higher education because of increasing demands for advanced training. This will lead to a need for a redefinition of the role and work of faculty in addition to their "roles as developers of curricula, planners of educational experiences, and managers and facilitators of student learning" (Lever, 1993, p. ix). Faculty must be a part of the redefinition process and must be rewarded if distance learning is to succeed (Dillon, 1989).

Thus, it is important to determine the nature of the instructional experience as viewed by the faculty member delivering the instruction and the students receiving the instruction. Researchers need to identify skills, knowledge and attitudes that are required by faculty members delivering instruction through distance learning.

Method

This study is a multi-method, multiple perspective approach to issues related to distance learning. Both

---

Neelam Kher-Durlabhji is currently serving as an Associate Professor in the Division of Education at Northwestern State University. Lorna J. Lacina-Gifford is serving as an Associate Professor in the Division of Education at Northwestern State University. Correspondence regarding this article should be sent to Neelam Kher-Durlabhji, Division of Education, Northwestern State University, Natchitoches, LA 71479 or by e-mail to kher@alpha.nsula.edu.

qualitative and quantitative paradigms are incorporated in the research design.

It focuses on the experiences of three faculty members who for the purposes of this study will be called Tom, Jane, and Susan. They all taught core courses in a Graduate Education Program at a small southern state supported university. Tom had more than twenty-five years of experience in the public schools in various teaching and administrative capacities. He joined higher education five years ago and taught graduate level education courses in Administration and Supervision. Jane had been in the teacher preparation program for over twelve years. She taught courses related to teacher strategies in the undergraduate preparation program and also taught graduate courses in Curriculum and Educational Philosophy and Leadership. Susan taught Educational Psychology courses in the undergraduate teacher preparation program and graduate level courses in Research, Statistics, Measurement and Learning. She had been teaching at the university level for six years.

The study also focuses on students in these classes who were either part of the studio or the remote site classes. All students were enrolled in a graduate degree program offered by the Division of Education. Of the students enrolled 60% were males. Less than 10% of the students were classified as ethnic minorities.

It is important to determine how students who experience distance learning evaluate the experience and if there is a difference in evaluation of the learning experience as a function of the site at which students receive the instruction.

Data Source

Data in the form of journals and extensive logs of the teaching experience, in depth interviews with the faculty, student evaluation of instruction, and video tapes of all aired classes were obtained. These data were used to elicit faculty reflections and interpretation of the instructional process. In depth interviews of the faculty members were conducted and transcribed by a graduate assistant who was trained in the techniques of qualitative interviewing (Guba & Lincoln, 1981; Merriam, 1988; Patton, 1990). Patton (1990) suggests that the best way to ensure reliability in interviewing is through adequate training. The researchers in this study provided extensive training to the graduate assistant who conducted the interviews.

Students' perspectives on the instruction is based on end of semester course evaluations. This end of semester evaluation instrument was developed by the researchers. The standard end of the semester instrument used by the university to ascertain instructional quality was deemed inappropriate because of the unique features of the distance teaching-learning environment. The instrument was reviewed by five experts in the field of research and measurement and revised based on the feedback. One hundred and seventy-one students provided the evaluation data. Of these 171 students, 67 received instruction in the studio setting and 104 were at 7 remote sites. The students were enrolled in the three graduate classes being taught via distance learning. No student took more than one distance learning course, thus no student reponded twice to the evaluation. These evaluations included a 15 item Likert-type scale focusing on students' opinions regarding distance learning compared to regular classrooms in terms of: existence of distractors; extent of attention; and ability to interact, ask questions, and relate to classmates.

Results

The first subsection of results focuses on the faculty perspective. The second subsection focuses on the data collected from the student perspective.

Faculty Perspective

Faculty reflections were gleaned from journals, logs of teaching experience, and extensive interviews. Based on the data, the following themes were identified: a) nature and description of courses being taught, b) preparation for teaching by satellite, c) impressions of being "live" on camera, d) modifications to the instructional process because of alternate delivery, and e) evaluation of the satellite teaching experience.

Courses being taught. Tom taught a graduate level course in School Law. In the studio setting there were 18 students and the other 39 students were at 7 remote sites in the state. Susan taught a required course in Educational Research. In this class 17 students were part of the studio classroom and 41 were at remote sites. Jane taught Educational Philosophy and Leadership with 18 students in the studio and 48 at remote sites.

In the traditional delivery system, the School Law was taught using the Case method, the Research course was primarily lecture, and the Philosophy and Leadership course had a large group participation component.

Preparedness for teaching. Jane had participated in a previous pilot satellite teaching project, Susan had evaluated some satellite courses, and Tom had no prior experience in the delivery system. As a part of the inservice, all three received a handbook outlining technical details, a two hour hands-on introduction to the studio facilities, and a brief ten minute practice session on camera. All three felt inadequate and taken aback as they faced class for the first time.

First impressions. Jane categorically stated, "I didn't like it. . . . there were just so many problems and a lack of control. . . it was difficult to concentrate on my lesson with commands being given in my earphone. . . . it was frustrating, just the lack of control and the feeling of being disorganized."

Tom said: "I felt scared to death. . . the live on camera was in one word frightening. I was not prepared to mentally focus on my subject matter without the focus on the delivery of the satellite instruction itself. It is one thing to interview and have your interview edited. But when you are on camera for two hours of instruction trying to focus on objectives, make it flow smoothly, and get interaction from the students, it becomes a different ball game. . . . the most important thing in your two hours is to manipulate the technology and be comfortable with it. The cameras took precedence over the content that I was trying to deliver."

Susan found the camera "quite intrusive" and the knowledge that the monitor was "right there and I could see myself teaching and that was very distracting . . . it put a distance between me and the students in the studio. I was conscious of being really worried. In addition to the camera you had to get used to camera people making all kinds of hand signals . . . so it was hard to keep my brain focused on the course and had to process all this other information and react intelligently."

Instructional modification. Modification that emerged from the experience with satellite teaching included the emphasis on several instructional modalities simultaneously to capture and sustain student attention (visual, auditory, kinesthetic); novel ways to involve students from the studio class and also the site students; and a high level of structure and preparation of instructional material to maintain flow and continuity of the delivery of content.

Susan summed it up saying, "One thing I had to modify a lot was that I couldn't be as impromptu as I wanted to be. . . . letting the class develop through the process of interaction. I had to be much more structured . . . much more prepared . . . to the point of knowing what jokes I was going to say at which point. This took a lot of prior planning because there was the feeling you wouldn't have the time to think once you were out there. . . . So the preparation had to be much more finely tuned going in and a great deal of the spontaneity was lost. The other thing I had to think about much more was providing visual stimuli.

Another thing had to do with motivating students to participate in the class and I had to consider it much more. . . . they needed to feel like part of the crowd so part way I decided to use captions (like those appearing on the Arsenio Hall Show) with student faces, twice during each class. Students enjoyed that and started contributing captions."

Handouts and other materials had to be mailed to the students at the remote sites a week in advance to make them available for discussion for class the following week. At the remote sites there was a facilitator whose role was to set up the TV system, distribute handouts, administer, proctor and mail back exams, etc. The facilitators were provided written instructions regarding their roles, duties, and expectations.

Evaluation of the satellite teaching experience. All three faculty members felt that they would be willing to teach another course using the satellite delivery system. However, the technological aspects of the instructional process were perceived to be out of their control and an undue influence on the course evaluations. The various arms of the instructional delivery system needed to work in sync to produce a coherent class. The negative feelings related to the teaching were closely tied to the experience of feeling out of control regarding various aspects of the delivery system.

Jane's feeling on teaching through distance learning encapsulated the sentiments of all three faculty members, ". . . I would not want to teach under the same conditions. I would only want to teach it if it was a different type of course. If it is a group discussion course, like a philosophy course, this is just not the best way to do it. If it was an information type class, one where you show examples of something, then I would consider it. But at this point I don't want to do another distance learning course under the same conditions."

Student Perspective

Descriptive statistics and *t*-tests were conducted to determine the pattern of responses in the studio and site groups (see Table 1) and to identify any significant differences that might exist between the groups (see Table 2). The *t*-test values and the probabilities associated with them are based on the Bonferroni correction. The Bonferroni correction was deemed appropriate because multiple *t*-tests were conducted (Christensen, 1996; Howell, 1982). Data from the three classes were pooled together because preliminary analysis did not reveal any differences between classes.

Table 1
Studio and Site Based Students' Reactions to Distance Learning

Items	Location	Percent Agree	Percent Neutral	Percent Disagree
I feel comfortable stating questions/concerns to instructor	Studio	34.4	19.4	46.3
	Site	59.6	4.8	29.8
I would take another distance learning class	Studio	41.8	13.4	10.5
	Site	72.2	9.6	12.6
Pace of instruction similar to regular class	Studio	12.4	13.4	29.8
	Site	32.6	21.2	40.3
Distance learning is less distracting than regular classes	Studio	16.4	16.4	37.1
	Site	31.8	27.9	34.6
I can pay greater attention in a distance learning class	Studio	22.4	19.4	58.3
	Site	36.6	24.0	33.6

Note: Studio *n*=67; Site *n*=104

There were significant differences in the responses of students by location. Students at the remote sites were much more satisfied with the turnaround time or feedback on assignments than the students attending the studio class. The students at the remote sites were also much more comfortable with note-taking and felt that they could use the same learning strategies as in the regular classroom. Students at the remote locations were much more likely to recommend a distance learning course to their friends. These data are presented in Table 2.

Table 2
Studio vs Site Students: Results of t-tests

Item	*N* of cases		Mean	*SD*	*t*	*p*
Feedback takes about the same time	Studio	67	4.82	3.22	3.06	.05*
	Site	104	3.50	2.01		
Note taking is just as easy	Studio	67	5.28	2.90	2.92	.05*
	Site	104	4.11	1.63		
Easy to get clarification	Studio	67	5.26	2.91	4.63	.001*
	Site	104	3.50	1.77		
Prefer regular course	Studio	67	5.56	2.65	5.47	.001*
	Site	104	3.59	1.79		
Used the same learning strategies	Studio	67	5.61	2.61	3.31	.001*
	Site	104	4.45	1.57		
Recommend distance learning courses	Studio	67	5.37	2.82	3.50	.001*
	Site	104	4.04	1.70		

They had concerns about the quality of the audio reception but not about the quality of video reception. Contrary to our expectations, the students did not feel that the physical absence of the teacher from the class made it easy for them to lose interest in the class.

In contrast, the studio class found the experience to be a greater distraction and did not feel that they could pay attention to the instruction as well. Because of classroom set-up, students in the studio class felt that they were missing out on classroom interaction with their peers. Students in the studio class felt self-conscious and found the pace of presentation to be much faster than in a regular classroom.

Conclusion

Results from this study suggest that students who are part of the studio experience view the distance learning experience quite differently than students who are at the remote sites. Students in the studio seem to feel like they are being "put on the spot." The environment in which the instruction is delivered seems contrived, with increased distance between the students and the instructor. The presence of the camera and attendant equipment, the possibility of the camera panning unexpectedly on the students, and changes in the seating arrangements all contribute toward making the experience less than optimal for the studio class.

The students at the remote sites seem to appreciate the distance learning because the instruction is now available "at their door step." Our findings indicate that students do not avail themselves of the interactive features of this instructional delivery system, are satisfied with the timelines of feedback they receive through the mail, and do not seem to need the physical presence of the instructor to maintain their attention.

Clearly, the experience of distance learning affects students in the studio setting in quite different ways than students at remote sites. To ensure the success of this alternative instructional delivery system, the instructional designers will need to address the concerns of the studio class. Ways to alleviate self-consciousness, anxiety about being on the camera and problems related to peer-interaction/instructor-student interaction will need to be addressed.

As universities try to meet the needs of a diverse population, they are more likely to explore alternative delivery systems of instruction to attract students who may be unable or unwilling to come to the main campus to take courses. Faculty are then required to modify their instructional strategies to better serve the needs of the students in the studio and at distant sites.

This study identifies a number of issues that will require research and debate as distance learning becomes more prevalent:

a) What factors determine the suitability of courses for alternate delivery systems? What modifications to traditional delivery systems will have to be made? How will this impact the quality of instruction?

b) How can faculty be trained to improve on camera delivery, and what new skills will they need?

c) How can the studio context be modified to remove or at least reduce the feeling of loss of control and alienation?

d) How can student evaluations of instruction be suitably modified to focus on aspects in the instructor's control?

In summary, distance learning can offer a powerful instructional alternative to traditional classroom based instruction. To ensure the effectiveness and quality of instruction in distance formats, faculty need greater training, support, and resources. Student feedback would be useful in addressing specific needs and concerns as they pertain to distance learning.

References

Barker, B. O., & Hall, R. F. (1993, October). *A national survey of distance education uses in rural school districts of 300 students or less.* Paper presented at the annual conference of the National Rural Education Association, Burlington, VT.

Bruder, I. (1991). Distance learning: Bridging education gaps with technology. *Electronic Learning, 11*(3), 20-23, 27-28.

Christensen, R. (1996). *Analysis of variance, design, and regression: Applied statistical methods.* London: Chapman & Hall.

Dillon, C. (1989). Faculty rewards and instructional telecommunications: A view from the telecourse faculty. *The American Journal of Distance Education, 3*(2), 35-43.

Eastman, D. V. (1994). Adult distance study through computer conferencing. *Distance-Education, 3*(2), 35-43.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation.* San Francisco: Josey Bass.

Howell, D. C. (1982). *Statistical methods for psychology.* Boxton: Duxbury Press.

Hyatt, S. Y. (1992, May). *Developing and managing a multi-modal distance learning program in the two-year college.* Paper presented at the Annual International Conference of the National Institute for Staff and Organizational Development on Teaching Excellence and Conference of Administrators, Austin, TX.

Lacina-Gifford, L. J., & Kher-Durlabhji, N. (1995, November). *Teaching a class by internet: A case study.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Lever, J. C. (1993). *Distance education resource guide: Computer-based distance education in community colleges.* (ERIC Document Reproduction Service No. ED 356 022)

Merriam, S. B. (1988). *Case study research in education: A qualitative approach.* San Francisco: Josey Bass.

Mizell, A. P., & Carl, D. R. (1994). Inter-Institutional cooperation in distance learning. *T.H.E. Journal, 21*(10), 91-93.

Ostendorf, V. A. (1989). *What every principal, teacher, and school board member should know about distance education.* Littleton, CO: Virginia A. Ostendorf, Inc.

Patton, M. Q. (1990). *Qualitative evaluation methods.* Newbury Park, CA: Sage.

Schrum, L. (1992, April). *Information age innovation: A case study of online professional development.* Paper presented at the annual conference of the American Educational Research Association, San Francisco, CA.

Speth, C. (1990, October). *Distance secondary education by satellite: An emerging research agenda - A review of the literature.* Paper presented at the Learning by Satellite Conference, Stillwater, OK.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 41-47

The Demise of The Georgia Teacher Performance Assessment Instrument

Dixie McGinty
Georgia State University

The Teacher Performance Assessment Instrument (TPAI) was used to assess the performance of beginning teachers in Georgia from 1980 to 1990. In this paper, the TPAI is treated from a political perspective, focusing in particular on the factors that led to its abolishment by the Georgia legislature in 1990. The case of Lee Ann Kitchens, a teacher who sued the State after losing her certificate because of failure on the TPAI, is examined in detail. Though it is tempting to view the phasing out of the TPAI as a direct result of the Kitchens case, this author suggests that the phasing out of the TPAI was actually the result of a number of different, though interrelated, factors.

During the mid-1970s, growing public skepticism about the quality of American education led many states to adopt accountability measures such as minimum competency standards. By 1978, minimum competency standards for students were in place in 33 states and were under consideration in most others. Public attention then turned to the issue of teacher competency. A new teacher certification testing movement arose in the South and eventually spread to other parts of the U.S. States began to implement teacher testing in a variety of forms, including paper-and-pencil certification tests, admissions testing for teacher education programs, and classroom observations.

In 1980, Georgia became the first state in the nation to require an on-the-job performance assessment for certification of beginning teachers. The Teacher Performance Assessment Instrument (TPAI) was a mammoth measurement tool that took more than four years and $5,000,000 to develop (B. Avera, Grants Accounting Supervisor for Georgia Department of Education, personal communication, March 18, 1993). In 1990, ten years after the TPAI was first implemented and five years after the final version was in place, the Georgia legislature voted to abolish it. In this paper, the author treats the TPAI from a political perspective, with primary emphasis on the factors that led to its abolishment. The phasing out of the TPAI provides some interesting insights into problems involved in teacher assessment.

Satisfactory performance on the TPAI was mandated for beginning teachers in May 1980. Three components were part of this measure. First, each teacher was to prepare a portfolio of lesson plans and materials. Second, an interview was conducted with the teacher to discuss the materials in the portfolio. Candidates were interviewed by three evaluators simultaneously: a peer teacher, an administrator, and a representative from a Regional Assessment Center, a sub-unit of the State Department of Education. The third component was an in-class observation by the three evaluators on one occasion, during which teachers were rated on a broad list of competencies, such as "reinforces and encourages the efforts of learners," "provides feedback to learners," and "demonstrates warmth and friendliness." These competencies were subdivided into more specific categories called "indicators," which were further specified by "descriptors," which provided evaluators with precise descriptions of the behaviors the candidate was supposed to exhibit. Each new teacher was allowed up to three years and a maximum of six tries to complete satisfactorily the requirements (Georgia Department of Education [GDE], 1979).

Early Support for the TPAI

One might imagine that the very idea of such a stringent evaluation procedure would evoke the opposition of teachers from its very outset. Before the initial implementation of the TPAI, however, there is little published evidence that teachers were against it. On the contrary, many teachers seemed to think that stricter requirements, including an on-the-job assessment, would help to upgrade their profession, bringing them the greater respect and higher pay they so desperately wanted and needed. The hope that higher standards might lead to greater rewards can be traced at least as far back as the

---

Dixie McGinty recently completed her doctorate and is a temporary Assistant Professor, Department of Educational Policy Studies, Georgia State University and after June 15, 1997, will become Assistant Professor of educational research, Western Carolina University. Please address correspondence to Dixie McGinty, 2400 Burch Circle, NE, Atlanta, GA 30319 or by e-mail: epsdlmx@panther.gsu.edu.

early seventies, when articles expressing it began to appear in the publications of the Georgia Association of Educators (GAE). In several other states, as well, evidence suggests that many teachers welcomed the idea of an on-the-job assessment to serve a gatekeeping function for the profession. For example, in a survey of Louisiana educators conducted two years prior to initial implementation, Chauvin and Ellett (1994) found widespread support for a similar teacher evaluation program.

By the end of the 1970s, the GAE had taken an official position on the TPAI issue: They would support the evaluation of beginning teachers, provided it was "properly administered" ("Proposed Resolutions," 1980, p.9). Even in the mid-eighties, when the instrument had been in use for five years, the GAE continued to support it, at least in principle. In a 1985 presentation to the State Board of Education, the GAE's executive director reiterated that "the GAE has been, and continues to be, one of the strongest advocates of performance-based teacher evaluation in the state" (Williams, 1985, p.3).

The lack of early organized teacher opposition to the Georgia TPAI may have meant that many teachers were actually in favor of it; however, it is also possible that the TPAI issue was simply overshadowed by other concerns that seemed more pressing at the time. For example, one issue that was very important on the GAE's agenda during the mid-eighties was the Georgia Teacher Certification Test (TCT), a subject-matter test which was a requirement for initial certification to teach. The anger and misgivings of many teachers regarding the TCT requirement won it a place in the *GAE Update* headlines month after month. Finally, the GAE filed suit against the State, claiming that the TCT was racially biased. In a sense, GAE opposition to the TCT may have actually created support for the TPAI, which many people viewed as a much better indicator of effective teaching than a paper-and-pencil test could ever be ("Testing," 1984).

Another issue on the forefront of GAE activity during the eighties was that of merit pay for teachers. As initially conceived, a merit pay plan would be based on an evaluation system. Though they still had not voiced opposition to the TPAI as such, GAE members strongly objected to the use of the TPAI to determine merit pay. Their fears that it might be used in this way were quieted when the State Board established a task force to begin developing a new instrument specifically for this purpose. Members of the task force included, among others, both the president and the executive director of the GAE. Plans for the new instrument probably absorbed enough interest to divert the GAE from worrying about the TPAI, at least for a while ("Teacher Evaluation Instrument," 1986, p.1).

Legal Challenge: The Kitchens Case

The first significant challenge to the TPAI, and one that would ultimately prove disastrous, was posed in 1988. Lee Ann Kitchens, an elementary school teacher from rural Meriwether County, filed suit after having lost her teaching certificate because she had failed the TPAI six times (*Kitchens v. State Department of Education*, 1988). When the case was decided in Kitchens's favor in a Fulton County Superior Court, the State Department of Education appealed the decision to the Georgia Supreme Court, where the ruling was upheld *(State Department of Education v. Kitchens*, 1990). Because the Kitchens case illustrates some important general aspects of teacher assessment, I will describe it here in some detail.

Lee Ann Kitchens first attempted the TPAI in November of 1984, her first year of teaching. She passed the portfolio and interview sections easily; it was the in-class observation that gave her difficulty. Even on this part of the assessment, she was judged to be adequate in 12 of the 14 competencies. Her inadequacies were in Competency 12 ("Demonstrates enthusiasm for teaching and learning and the subject being taught") and Competency 13 ("Helps learners develop positive self-concepts"). After her first few tries, Kitchens was rated satisfactory on all the indicators within these competencies except "communicates personal enthusiasm," which was part of Competency 12, and "demonstrates warmth and friendliness," part of Competency 13 (Rowe, 1987).

After failing the observation several times, Kitchens sought assistance from the Regional Assessment Center, which she had been told offered staff development courses for unsuccessful TPAI candidates. Representatives from the Assessment Center told her there was nothing they could do for her. After all, they told her, "warmth" and "enthusiasm" were qualities that could not be taught. Kitchens then requested that her next observation session be videotaped so that she could actually see what she was doing wrong. This request was denied. The required post-evaluation conferences were of little value because, in Kitchens's words, "the person that came [for the conference] was never the same one that evaluated you" (L.A. Kitchens, personal communication, July 31, 1991).

In 1987, Kitchens failed the TPAI for the sixth time. Perhaps she did not "smile at learners or laugh or joke with them" (GDE, 1979, p.41), or "use the names of learners in a warm and friendly way" (Georgia Department of Education, 1979, p.41). Even more unthinkable, she may have neglected to "communicate enthusiasm with gestures to accentuate points" (Georgia

431

Department of Education, 1979, p.40). Whatever the specific reasons, Lee Ann Kitchens failed to meet the standards set by the State for minimal teaching competency. Her teaching certificate was revoked, and she lost her job.

Kitchens reports that it was her principal at Luthersville Elementary who encouraged her to challenge the decision. Initially, she appealed through the channels established by the State Department of Education, even though Regional Assessment Center Representatives had told her that "no one ever won" (L.A. Kitchens, personal communication, July 31, 1991), which is not surprising because appeals of the TPAI were heard by the same Department of Education division that administered it. At the hearing in November 1987, the Special Hearing Master was not sympathetic to Kitchens's plight. After all, he remarked, Kitchens had been evaluated by 14 different observers during the course of her six attempts to pass the TPAI. Further, he emphasized that the TPAI had been so thoroughly validated that "the possibility of misclassifying an individual is infinitesimal" ("Certification hearing," 1987). The decision against Kitchens was not reversed.

The following summer, GAE staff attorney Bensonetta Lane, to whom Kitchens had been referred for legal representation, filed with the Fulton County Superior Court a petition seeking reversal of the decision (*Kitchens v. State Department of Education*, 1988). Her argument centered around three issues that had not been resolved at the Department of Education hearing. First, she claimed that Kitchens had not received enough information about how to do well on the TPAI. In response to this, attorneys for the Department of Education promptly presented a statement that Kitchens had signed in 1984 certifying that she had been given an adequate orientation to the TPAI process. Lane's second argument was that Kitchens had been denied her right to due process; to this the State countered that "the finite number of opportunities to pass the TPAI does not constitute a lack of due process" *(Kitchens v. State Department of Education*, 1988). Finally, Lane submitted that the TPAI should be declared invalid because regulations governing it had not been drafted in accordance with the Georgia Administrative Procedure Act (GAPA). This act requires that proposed legislation be posted with the Secretary of State so that the public may comment on it before enactment (*Administrative Procedure Act*, 1964).

This last issue proved to be an enormous thorn in the side of both the Department of Education and the State Board. Attorneys for the State took the position that the GAPA did not apply to the State Department of Education, nor to the Board, because the act included wording that specifically excludes "any school, college..., or other such educational...institution" (*Administrative Procedure Act*, 1964). Fulton County Superior Court Judge Frank Eldridge, however, ruled that the Department of Education and the Board of Education were not educational institutions as such, but rather agencies of the type to which the GAPA was intended to apply.

But the State was not willing to accept defeat. In demanding a remand, State attorneys pointed out that the General Assembly had not required the State Board to comply with the GAPA in legislating the reforms of the Quality Basic Education (QBE) Act in 1985 and 1987. The court's ruling in the Kitchens case, they argued, thus rendered all the reforms of the QBE act invalid, which would lead to a disruption in the operation of schools throughout the state. Bensonetta Lane's response to this scenario was that "our system of justice cannot operate as if convenience [were]...a controlling factor" (Lane, 1988).

Eldridge stood firm in his decision in Kitchens's favor. The following year, the State Department of Education appealed the case before the Georgia Supreme Court. There the decision was upheld because the exclusion of educational institutions from the GAPA was "intended to apply only to those institutions which provide educational services directly, not to the administrative bodies through which they are regulated" (*State Department of Education v. Kitchens*, 1990). Lee Ann Kitchens, along with some 135 other former teachers (White, 1990), was invited to reapply for a teaching certificate. She was promptly rehired by the Meriwether County School System. The following January, the Georgia General Assembly passed legislation amending the GAPA so as to state explicitly that it applied to both the State Board and the Department of Education. The two agencies were given one year to redraft all their regulations to comply with the GAPA (H. R. Bill 1520, 1990).

The Kitchens case provides a wealth of insights into the problems associated with the performance assessment of teachers. First, it dramatically illustrates the difficulties inherent in attempting to measure qualities like "warmth" and "enthusiasm," and--perhaps even more important--the necessity of ensuring that candidates know what they can do to improve. Second, attention is called to the overly defensive posture of the State, which, by not allowing observations to be videotaped, created a relationship with candidates that was far more adversarial than cooperative. Third, a number of issues are raised related to the local administration of such an instrument.

For example, was it State policy that the post-observation conference be conducted by someone other than the person who had actually made the observation? Or was it simply that lax administration guidelines allowed the Regional Assessment Center staff to avoid unpleasant confrontations with the candidates they had evaluated by sending a different staff member to discuss the evaluation? In addition, the case calls into question the motivation behind many local decisions. Recall that Kitchens's principal encouraged her to appeal the decision against her; Kitchens also reports that he was "shocked" when she did not pass. Yet, in the legal file on the Kitchens case, handwritten notes indicate that the principal told Assessment Center representatives, after Kitchens's fifth attempt, that she still had "a long way to go." One wonders how he really felt about her performance.

The Demise of the TPAI

In the aftermath of the Kitchens case, opposition to the TPAI grew. The publicity generated by the lawsuit had brought the TPAI into the public eye, and, in doing so, had opened doors to new opposition. Teachers had varying complaints, many of which were unrelated to those voiced in the Kitchens case. They charged, for example, that the portfolio requirement was far too time-consuming for a beginning teacher, and that their portfolios were being judged on the basis of form rather than substance ("Teachers Blame QBE," 1988).

The GAE had gradually become more forceful in its opposition to the TPAI, although its complaints were still centered around issues of administration and not on the instrument itself. GAE members argued, for example, that TPAI orientations were inadequate, that candidates were seldom informed of the appeals process, and that the skills of the evaluators varied from region to region ("1988 Legislative Priorities," 1987). With the Department of Education's position weakened somewhat by the GAPA controversy, GAE lawyers were able to win for their clients additional chances to try to pass the TPAI. In its official list of legislative priorities for 1990, the GAE declared its intention to "support legislation to correct the professional inequities and abuses currently being practiced by the State Department of Education in the TPAI" ("1990 GAE Legislative Priorities," 1989, p.6).

In the 1990 session of the Georgia legislature, Senator Sallie Newbill of Roswell introduced a bill to revise the requirements for beginning teacher certification and eliminate the use of the TPAI (H. R. Bill 439, 1990). The original bill was rejected by the House Education Committee, but a similar bill retaining much of Newbill's original wording passed both houses unanimously. According to the bill, an in-class observation would still be required for beginning teachers, but it would be done using the Georgia Teacher Observation Instrument (GTOI), which was then being used for annual evaluations of experienced teachers (S. Bill 1212, 1990).

Newbill's motivation for introducing the new legislation is not entirely clear, though she appears to perceive herself as having done so in response to the immediate needs of her constituents. She reports that her interest in the TPAI issue was aroused through her regular attendance at meetings of a North Fulton County association of parents and teachers (S. Newbill, personal communication, August 7, 1991). James Fox, superintendent of the Fulton County School System, had long been opposed to "objective" evaluation instruments like the TPAI. In his view, the TPAI was "designed in a sterile atmosphere to serve university purposes" (J. Fox, personal communication, August 12, 1991). Fox worked closely with Newbill to develop the proposed legislation and to build support for it with the Atlanta newspapers, the governor, and others. In the March 1990 issue of its *Update*, the GAE congratulated itself on its successful lobbying to abolish the TPAI ("GAE Lobbies," 1990). Senator Newbill, however, claims to have been unaware that any such lobbying was going on. If the GAE was lobbying, she said, she "never saw them during that entire session" (S. Newbill, personal communication, August 7, 1991).

The Georgia Teacher Observation Instrument (GTOI), which was now to be used to evaluate beginning teachers, differed from the TPAI in several ways. For one thing, it was simply an observation instrument and did not include a portfolio component. Compared to the observation portion of the TPAI, the GTOI was much simpler and much less specific, leaving more leeway to the evaluators. The final version of the TPAI included a total of 92 behaviors. The observer was to determine, using a checklist, whether the teacher had or had not demonstrated each of them. The GTOI, in contrast, required the evaluator to make decisions about candidates on a mere 15 "dimensions." Though the instrument included explanatory text defining each dimension, it was essentially up to the evaluator to determine whether the candidate performed adequately on each. Most of the space on the actual observation form was to be used for written comments by the evaluator.

The GTOI was, in many ways, more palatable to teachers and administrators than was the TPAI, but its results were less quantitative and therefore less useful in justifying difficult decisions such as the revoking of teaching certificates. In the spring of 1990, State Superintendent Werner Rogers, possibly motivated by misgivings about the use of a less precise instrument for such purposes, recommended to the State School Board that on-the-job assessment be dropped as a requirement for initial certification. The rationale Rogers presented to the Board was that it was no longer necessary to assess new teachers because all teachers were now being evaluated annually using the GTOI. In a June meeting of the Board, the recommendation was approved. The Board also directed the Department of Education to broaden the Teacher Certification Test by adding new components to test writing skills and teaching methods. It also stipulated that the Department explore reestablishment of a performance-based assessment requirement for initial certification in the future ("State Board of Education Discontinues," 1990). Within the next five years, however, the idea of decentralized control of schools had gained so much momentum in the state that reinstating such a requirement for certification would have been virtually unthinkable.

The phasing out of the TPAI is best viewed as the result of a number of different, though interrelated, factors. It may be tempting to view it as a result of *Kitchens v. State Department of Education*, yet those who introduced legislation to abolish it appear to have been motivated by reasons quite unrelated to the case. What seems to have happened is that the publicity given to the TPAI as a result of the Kitchens case brought its negative points very much into public consciousness, thereby coaxing already-simmering opposition into a full boil.

In its zeal to implement an evaluation process that teachers would find palatable and, above all, litigation-proof, the State may have defeated its own purposes. For example, refusal to allow videotaping "benefited" the State by depriving potential plaintiffs of actual recordings they could use to challenge their evaluations. But Lee Ann Kitchens reports that not being allowed to videotape her session was one of her greatest frustrations with the process, and one of the main reasons why she sued (L.A. Kitchen, personal communication, July 31, 1991). The high degree of specificity of the TPAI can also be viewed as the State's attempt to thwart potential challenges. Yet there are substantial difficulties in attempting to parse the act of teaching into minute bits. Perhaps it was inevitable that teachers would eventually begin to feel that the very instrument they had hoped would upgrade their profession had actually degraded it.

Discussion

By 1988, classroom assessment programs for beginning teachers were also in place in Florida, Kentucky, Louisiana, North Carolina, Oklahoma, South Carolina, and Virginia. Some of these have survived to the present, while others have not. Whether or not such a program can last depends not only on political factors specific to each state, but also on variables such as the degree of support provided to candidates, fairness of administration, and the type of instrument.

Ostensibly, teacher assessment programs in most states served the dual purposes of providing support to the beginning teacher and determining whether the candidate would be granted regular certification; however, as Rudner (1987) notes, much of the expected support activity simply did not take place. In Georgia, the assistance offered to teachers was minimal, and the support function was clearly subordinate to the gatekeeping function. Oklahoma's Entry-Year Assistance Program, in contrast, emphasized the support function. Each entry-level teacher was assigned to a teacher consultant, who was required to work with the candidate at least 120 days and spend at least 72 hours in consultation or observation (Hopkins, Elsner, & Shreck, 1983). Teacher consultants received an extra $500 per year above their regular salary. The Oklahoma program is still in place today, 15 years after its initial implementation.

Georgia was not the only state in which complaints about administration procedures may have contributed to the failure of a teacher assessment program. Educators' attitudes toward the Louisiana Teaching Internship Program (LTIP), for example, were investigated in a series of surveys conducted over a four-year period (Chauvin & Ellett, 1991, 1994; Chauvin, Evans, & Ellett, 1992). Results revealed "a significant disparity between perceptions held prior to statewide implementation regarding what the programs 'would be like' and how they 'really are'" (Chauvin & Ellett, 1994, p.10). Respondents cited problems with inconsistent implementation, unclear communication, inadequately trained assessors, and inadequate support at the school and district levels. At the end of its first year of statewide implementation, the LTIP was suspended by the Louisiana state legislature.

Even with proper administration procedures and adequate support services for candidates, teacher performance assessment is problematic. What type of instrument should be used to measure teacher effectiveness? The instruments used by the states mentioned above were, for the most part, quite similar to the Georgia TPAI; several were actually modeled on Georgia's instrument. More recently, especially from the ideological perspective of the National Board for Professional Teaching Standards (NBPTS), it has been argued that highly specific, behavior-based instruments cannot adequately capture the complexity of the act of teaching.

If one believes, as I do, that some form of teacher evaluation is probably necessary, one is faced with a very real dilemma. Teaching is an art. Any instrument used to evaluate it will fall short of its goal. A highly quantitative instrument like the TPAI will not only fail in its goal of objectivity; it will also, in aspiring to this goal, reduce the teaching process to a list of mechanical behaviors, robbing it of its richness and creativity. A more open-ended instrument, on the other hand, is politically risky, especially when adverse decisions must be made based on its results. This problem is discussed at length by Delandshere and Petrosky (1994), who cast it in a post-structuralist framework.

Good teaching can be achieved through a vast, if not infinite, variety of approaches. Can assessment instruments be developed that allow for this variety without sacrificing the reliability that is so important for high-stakes uses in a litigation-happy society? Perhaps not. No easy solution exists to the problem, and any progress toward one will require the continuing collaboration of policymakers, educators, and the measurement community.

References

1988 legislative priorities. (1987, May). *GAE Update* [publication of the Georgia Association of Educators], p. 6.

1990 GAE legislative priorities. (1989, June). *GAE Update* [publication of the Georgia Association of Educators], p. 6.

Administrative Procedure Act, Ga. Code Ann. §50-13-2 (1964).

Certification hearing appeal of Lee Ann Kitchens to the Georgia Department of Education (1987, November 17) [on file under *Kitchens v. State Dept. of Educ.*, Fulton County Superior Court, Atlanta, Georgia].

Chauvin, S. W., & Ellett, C. D. (1991). *Reflecting on teaching and learning: A study of school-level implementation of the STAR in nine schools.* Technical Report Number 18, Louisiana Teaching Internship and Teacher Evaluation Projects, College of Education, Louisiana State University, Baton Rouge, Louisiana.

Chauvin, S. W., & Ellett, C. D. (1994, April). *"The morning after": Year IV study of an evaluation program designed to replace lifetime teacher certification with a renewable credential.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Chauvin, S. W., Evans, R. L., & Ellett, C. D. (1992, April). *Replacing lifetime certification with a renewable credential: Phase III of a study of Louisiana educators' perceptions about the Louisiana Teaching Internship and Statewide Teacher Evaluation Programs.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge: Performance assessment. *Educational Researcher, 23*(5), 11-18.

GAE lobbies successfully to abolish TPAI. (1990, March-April). *GAE Update* [publication of the Georgia Association of Educators], p. 4C.

Georgia Department of Education (1979). *Teacher performance assessment instruments.* Athens, Georgia: University of Georgia Performance Assessment Laboratory.

Hopkins, S., Elsner, K., & Shreck, G. (1983, February). *Entry-year assistance committee: A support system for beginning teachers.* Paper presented at the Annual Meeting of the Association of Teacher Educators, Orlando, Florida.

H. R. Bill 439, 1990 Georgia General Assembly.

H. R. Bill 1520, 1990 Ga. Laws 1520.

Kitchens v. State Dept. of Educ., No. D-54773 (Fulton County Sup. Ct. 1988).

Lane, B. (1988, August 22). Brief in opposition to defendant's motion for remand or, in the alternative, to set aside judgment [on file under *Kitchens v. State Dept. of Educ.*, Fulton County Superior Court, Atlanta, Georgia].

Proposed resolutions--1990 convention. (1980, March). *GAE Update* [publication of the Georgia Association of Educators], p. 9.

Rowe, E. (1987, April 9). Memorandum to Lester Solomon [on file under *Kitchens v. State Dept. of Educ.*, Fulton County Superior Court, Atlanta, Georgia].

Rudner, L. (1987). *What's happening in teacher testing: An analysis of state teacher testing programs.* Washington: Office of Educational Research and Improvement.

S. Bill 1212, 1990 Ga. Laws 1212.

State board of education discontinues performance-based assessments. (1990, June). *GAE Update* [publication of the Georgia Association of Educators], supplementary insert page.

State Dept. of Educ. v. Kitchens, 259 Ga. 791, 193 Ga. App. (1990).

Teacher evaluation instrument and procedure under development. (1986, October). *GAE Update* [publication of the Georgia Association of Educators], p. 1.

Teachers blame QBE for paperwork avalanche. (1988, January 24). *The Atlanta Constitution*, p. B1.

Testing for veteran teachers? (1984, September). *GAE Update* [publication of the Georgia Association of Educators], p. A3.

White, B. (1990, January 27). Court gives new hope to teachers who failed state reviews. *The Atlanta Constitution*, p. D1.

Williams, J. (1985, June-July). Use of the TCT to test veteran educators for recertification and performance evaluation as a pre-requisite to awarding annual experience increments. *GAE Update* [publication of the Georgia Association of Educators], p. 3.

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 49-59

Responses That May Indicate Nonattending Behaviors in Three Self-Administered Educational Surveys

J. Jackson Barnette
University of Alabama

Because of their widespread use, it is critical that surveys used to collect data for research and evaluation purposes be as valid and reliable as possible. Identifying and removing sources of measurement error are always important. One source of such error is using data from respondents who, for whatever reason, do not attend to the survey items. The incidence of three types of respondent behaviors that are consistent with nonattending respondents in three self-administered surveys was examined. The three behaviors were: missing items, item response patterns, and attending to reverse-worded items. The three surveys included: one on attitude toward elements of children's literature taken by $n=1240$ *elementary students, one on attitude toward classroom questioning taken by* $n=3541$ *students in grades five through twelve, and one on attitude toward classroom questioning taken by* $n=2688$ *teachers. While there were relatively few respondents who left out more than 10% of the items, about 23% of the respondents left at least one item blank. About 14% of the respondents in the children's literature survey provided patterns of responses which were consistent with nonattending behaviors. When comparing the negative-positive response distributions of item responses between direct-worded items and reverse-worded, reverse-scored items, 31% of the students and 26% of the teachers in the classroom questioning surveys had significantly different distributions at* $p < .05$.

A great deal of educational research and evaluation is based on the use of self-administered attitude surveys, those where the respondent marks the survey form or optical scanning sheet. These are often administered to a voluntary or involuntary, somewhat captive, intact group. In such situations, there is often concern about whether respondents are attending to the items in an attentive, thoughtful manner. A respondent who, for whatever reason, either purposely or not purposely, does not attend to answering the items in an attentive, thoughtful manner is considered a nonattending respondent. Such behavior could be manifested in several ways, among them: leaving items missing, providing responses without reading or thinking about items, and failing, if they are present, to attend to reverse-worded items. Nonattending respondents introduce measurement error into the data set, and thus it is important to identify such behaviors, examine their effects, and determine if respondents should be deleted from the data.

The purpose of this paper is to examine results from three different surveys relative to the incidence of behaviors that are consistent with nonattending respondents:

1. Missing items throughout and at end of survey,
2. Item response patterns that are consistent with lack of attentiveness, and
3. Failure to respond to reverse-worded items.

Related Literature

The issue of error or bias associated with attitude assessment has been discussed for the past several decades. Cronbach (1970, pp. 495-499) discussed two behaviors which bias responses, those of faking and acquiescence. Faking behavior is characterized by a respondent consciously providing invalid information such as in providing self-enhancing, self-degrading, or socially desirable responses. Acquiescence relates to the tendency to answer in certain ways such as tending to be positive or negative in responding to Likert-type items. Hopkins, Stanley, and Hopkins (1990, p. 309) presented four basic types of problems in measuring attitude: fakability, self-deception, semantic problems, and criterion inadequacy.

Nunnally (1967, pp. 612-622) indicated that some respondents have an extreme-response tendency, the differential tendency to mark extremes on a scale; some have a deviant-response tendency, the tendency to mark responses that clearly deviate from the rest of the group. If such responses are thoughtful and, from the viewpoint of the respondent, representative of true opinions, then they should not be considered nonattending or spurious. However, if respondents mark extremes or deviate from

---

An earlier version of this manuscript was presented at the Annual Meeting of the Mid-South Educational Research Association, November 9, 1995. J. Jackson Barnette is Professor of Educational Research in the College of Education at The University of Alabama. Correspondence regarding this paper should be sent to J. Jackson Barnette, 2428 Brandon Parkway, Tuscaloosa, AL 35406 or by e-mail at jbarnett@dbtech.net.

the group because of reasons not related to their opinions, then they would be considered to be nonattending respondents. Nunnally also discusses the problems of carelessness and confusion. These are more likely to be similar to what may be referred to as nonattending respondents. Respondents who are careless or confused, yet are forced, either directly or indirectly, to complete the survey are more likely to provide spurious or nonattending responses.

Lessler and Kalsbeek (1992) point out that "There is disagreement in the literature as to the nature of measurement or response variability. This disagreement centers on the nature of the process that generates measurement variability" (p. 277). They refer to the problem of individual response error where there is a difference between an individual observation and the true value for that individual.

In a review of several studies related to response accuracy, Wentland and Smith (1993) concluded that "there appears to be a high level of inaccuracy in survey responses" (p. 113). They identified 28 factors, each related to one or more of three general categories: inaccessibility of information to respondent, problems of communication, and motivational factors. They also report that in studies of whether the tendency to be untruthful in a survey is related more to personal characteristics, item content, or context characteristics, personal characteristics seem to be more influential. As Groves (1991) points out: "Different respondents have been found to provide data with different amounts of error, because of different cognitive abilities or differential motivation to answer the questions well" (p. 3).

One behavior which would indicate nonattending behavior is not answering items. Missing items, in particular those at the end of a survey, may indicate nonattending behavior. Of course there are different reasons this may happen including lack of understanding, reading difficulty, fatigue, lack of interest or motivation to participate, lack of time, and others. Regardless of the reason, such behavior, if prevalent, provides respondent bias because true attitude is not being assessed. While there have been several studies which examine different methods of replacing missing data (Huberty & Julian, 1995; Kaiser, 1990; Witta & Kaiser, 1991;), there is a lack of research on how often missing data occur and where in the data set they occur. Witta (1994) concluded that missing data in a large national survey were not missing in a random manner.

Goldsmith (1988) conducted research on the tendency of providing spurious responses or responses which were meaningless. In a study of claims made about being knowledgeable of product brands, 41% of the respondents reported they recognized one of two fictitious product brands and 17% claimed recognition of both products. One group identified as providing more suspect results was students. In another study (Goldsmith, 1989) where respondents were permitted to respond "don't know" and were told that some survey items were fictitious, the frequency of spurious response decreased, but not by very much. Goldsmith (1986) compared personality traits of respondents who provided fictitious responses with those who did not when asked to indicate awareness of genuine and bogus brand names. While some personality differences were observed, it was concluded that the tendency to provide false claims was more associated with inattention and an agreeing response style. In Goldsmith's research it is not possible to separate out those who purposely provided spurious responses as opposed to those who thought they were providing truthful answers. Perhaps only those who knowingly provided fictitious responses should be considered as providing spurious responses.

Respondent error could result from collection of data from different types of respondents. A study conducted by Marsh (1986) related to how elementary students respond to items with positive and negative orientation found that preadolescent students had difficulty discriminating between the directionally-oriented items and such ability was correlated with reading skills. Students with poorer reading skills were less able to respond appropriately to negatively worded items. Pilotte and Gamble (1990) concluded that when a mix of direct and reverse-worded item stems are used as compared with all direct-worded stems, the two sets of items do not define a single construct and the instrument is less reliable.

When there are a large number of respondents, it is assumed that a small proportion of nonattending respondents is not likely to have much affect on the commonly used statistics. In many situations this is the case. In the review of the literature, there were no sources found which indicate how prevalent such respondents might be in different situations. In addition, there were no definitive studies of the effects of different types of response patterns which may be used by nonattending respondents on the commonly used statistics associated with self-administered questionnaires. Barnette (1996), pointed out that different response patterns have different effects on Cronbach's alpha. For example, one pattern consistent with nonattending behavior would be when all responses are at the same extreme value when the typical pattern is in the middle of the scale. He found that when five percent of the respondents are randomly replaced with responses that are all at an extreme in a population where alpha was .70, alpha increased to .89. If five percent of the respondents were replaced with respondents who responded half at one extreme and half at the other extreme, alpha decreased to .66.

There are several possible reasons for a respondent to not attend to items on a survey. Among them would be, the respondent may: (1) not understand the directions or items; (2) lack the experience or knowledge to accurately respond to the items; (3) lack the motivation to accurately respond to the items; (4) purposely over exaggerate responses out of anger, frustration, or some other emotional condition related or unrelated to the questionnaire topic; (5) want to finish the task as quickly as possible; and (6) be or become fatigued in the process of completing the survey, particularly a long one. Regardless of the reason for nonattending, measurement error is increased in the data set.

As Bradburn and Sudman (1991) point out: "It is unrealistic to expect that most measurement errors will ultimately be eliminated. Optimistically, we should attempt to reduce those effects that can be reduced using the best survey methods and to understand and be able to measure the size of the effects that cannot be reduced" (p. 31).

Method

Subjects

Three rather large actual data sets are examined in this study. These were selected because they represented different respondent groups, they had large sizes, and they were available. Summary characteristics for each of these data sets are found in Table 1. Survey 1 was a survey of attitudes toward elements of children's literature. There were three forms of this survey: one was read to kindergarten and first graders and the students circled a very sad, sad, happy, or very happy face; one was written in very simple language for second and third graders, and the third was at a higher reading level and was given to fourth through sixth graders. All three forms had 25 parallel items, all of which were worded in the same direction with a high score, on the four-point scale representing a more positive attitude. Examples of the items on these surveys are: "Do you like books which surprise you or have a surprise ending?"; "Do you like stories in which a character must overcome something all by himself?"; "Do you like survival stories in which someone must struggle against nature?"; and "Do you like books that keep you guessing what will happen next?"

Data were collected from 1240 students from 59 different classrooms in 27 different schools in northern Alabama. Cronbach's alpha for this survey was .817, based on 922 respondents who answered every item. The overall mean rating was 2.831 and the item standard deviation was 1.129.

Table 1
Characteristics of Each Survey Data Set

	Survey 1 Grades K-6	Survey 2 Grades 5-12	Survey 3 Teachers
Sample size	1240	3541	2688
Number of items	25	57	50
Number of reverse worded items	none	14	19
Item mean	2.831	2.177*	2.007*
Item standard deviation	1.129	0.925*	0.792*
Alpha internal consistency	0.817	0.875*	0.875*
Alpha based on *n* of	922	2633	2207

*Based on reversed items for Surveys 2 and 3

Survey 2 was a survey of student attitudes toward classroom questioning. It had 57 items, each on a strongly disagree to strongly agree, four-point scale. Fourteen of the items were reverse-worded such that either "disagree" response was a positive response. Data were collected from 3541 students representing grades five through 12 from more than 15 school districts in five different states. After reverse scoring the 14 reverse-worded items, Cronbach's alpha for this survey was .875, based on 2633 respondents who answered every item; the overall mean rating was 2.177; and the item standard deviation was .925. Examples of items on this survey are: The questions asked in this class help me learn the class material better; I am often afraid that the questions I ask might be considered dumb questions; My teacher allows me enough time to think before I am required to give an answer; and My teacher uses our answers to think up new questions.

Survey 3 was a survey of teacher attitudes toward classroom questioning. It had 50 items, each on a strongly disagree to strongly agree, four-point scale. Nineteen of the items were reverse-worded such that either "disagree" response was a positive response. Data were collected from 2688 teachers representing elementary, middle, and secondary schools from more than 20 school districts in five different states. After reverse scoring the 19 reverse-worded items, Cronbach's alpha for this survey was .875, based on 2207 respondents who answered every item; the overall mean rating was 2.007; and the item standard deviation was .792. Examples of items on this survey are: "Effective classroom questioning improves student achievement"; "Good questions come naturally to most teachers"; "Pausing after a question (before calling on a student to answer) wastes class time"; and "It is important for students to hear only correct responses to questions during instruction."

Analysis

Missing Items. For all three data sets, the frequency and percent of missing items throughout the survey and frequency and percent of missing items at the end of the survey were determined. Proportions of missing items were compared across the three surveys using chi-square homogeneity of proportions tests.

Response Patterns. Item response patterns may indicate nonattending behaviors. For example, a respondent who marks the same response for every item may not be attending, especially if there is a relatively high degree of item response variability for the total group. Such respondents have low item variability and will be referred to as monotonic respondents. Item means could be at any point across the scale. A monotonic respondent who marked all at either the lowest or highest response set will be referred to as a mono-extreme while a respondent who marked all the same responses at a point close to the item mean for the total group will be referred to as a mono-middle respondent.

Another possibility of nonattending behavior would be marking half or a high proportion of each of the most extreme response possibilities. This respondent, which will be referred to as a checker-extreme, has very high item variability but an item mean around the middle of the scale.

Examination of the item mean and variability patterns may provide a basis for identification of potential nonattending respondents. Of course, in any fixed response set, such as with Likert items, there is a relationship between the item mean and item variance. The closer the mean gets to one of the extremes, the lower the maximum possible item variance. If all items are marked at an extreme there can be no item variance. Maximum item variance is possible when responses are in the middle of the scale. Even though the item mean and item variance are not independent, certain combinations reflect response patterns.

Figure 1 provides a matrix of patterns of z deviations of respondent means and standard deviations from the total group values. Entries in the cells represent different row and column combinations. Cut-off values of the normal distribution to separate the distributions into lowest 10%, next 20%, middle 40%, next 20%, and highest 10% were used to determine the categories of this matrix. Based on expected cell percentages, under an assumption of normal distributions of respondent item means and standard deviations, the expected percentages are computed and indicated in the cells. While most cells are not identified with any specific type of potential nonattending respondent, others are. Cell 11 is indicative of mono-extreme patterns at the very low end of the scale, cell 51 is indicative of mono-extreme at the very high end of the scale, and cell 31, and to a lesser extent cells 21 and 41, are indicative of mono-middle respondents. Cells 15, 25, and 35 have very high variability with a mean close to the over all item mean, a pattern similar to what would be expected of a checker-extreme.

z_s deviations → z_m deviations ↓	Very low SD <-1.28	Low SD -1.28 to <-0.52	Middle SD -0.52 to +0.52	High SD >+0.52 to +1.28	Very high SD >+1.28
Very low mean <-1.28	11 (.01)	12 (.02)	13 (.04)	14 (.02)	15 (.01)
Low mean -1.28 to <-0.52	21 (.02)	22 (.04)	23 (.08)	24 (.04)	25 (.02)
Middle mean -0.52 to +0.52	31 (.04)	32 (.08)	33 (.16)	34 (.08)	35 (.04)
High mean >+0.52 to +1.28	41 (.02)	42 (.04)	43 (.08)	44 (.04)	45 (.02)
Very high mean >+1.28	51 (.01)	52 (.02)	53 (.04)	54 (.02)	55 (.01)

(proportion expected under normal distribution assumption)

Figure 1. Matrix of z standard error deviations from item means and standard deviations

The frequency and percent of respondents falling into these 25 cells were determined for only the Survey 1 data set. Since Surveys 2 and 3 had reverse-worded items, this analysis would not be legitimate. Only respondents with 70% or more responses were considered. The respondent item mean was computed and converted to a *z* score based on the standard error of item mean for the entire group. Then the respondent item standard deviation was computed and converted to a *z* score based on the standard error of the item standard deviation.

Reverse-Worded Items. The third analysis compared the distributions of direct and reverse-worded items for Surveys 2 and 3. Since the items were very similar and the reverse-worded items were distributed throughout the survey, it is assumed that if the reverse-worded items were reverse-scored they should demonstrate a similar frequency distribution of responses compared with the frequency distribution of items that were not reverse-worded. For this analysis the reverse-worded items were reverse scored (4 to 1, 3 to 2, 2 to 3, and 1 to 4). The two disagree responses were collapsed and the two agree responses were collapsed to ensure adequate cell expected frequencies. The frequency distributions were compared using one-degree of freedom chi-square statistics. The probabilities were computed for each respondent. A significant chi-square would indicate a difference in the proportions of disagree and agree responses between the direct and reverse-worded items, a situation which would be consistent with nonattending behavior.

Results

Missing Responses

Table 2 presents the results for the missing item analysis for Survey 1, which had 25 items. Of the 1240 respondents, 922 (74.4%) answered all items, 14.4% had one missing item, 5.1% had two missing items, and 6.2% had more than ten percent if the items missing. Only 1.8% had item scores missing at the end of the survey and most of them had only the last item missing.

Missing item results for Survey 2 are presented in Table 3. Of the 3541 respondents, 74.4% provided responses for all 57 items, 16.2% had one missing item, 4.9% had two missing items, and 4.5% had three or more items missing. Only 2.3% had any items missing at the end of the survey. Slightly more than one percent (1.1%) had more than ten percent of the items missing at the end of the survey.

Table 2
Percent of Items Left Missing Across
and at End for Survey 1, 25 Items

Number missing	Percent missing	Across all items *f*	Across all items %	At end of survey *f*	At end of survey %
0	0.0	922	74.4	1218	98.2
1	4.0	178	14.4	14	1.1
2	8.0	63	5.1	0	0.0
>2	more than 10	77	6.2	8	0.6

n= 1240

Table 3
Percent of Items Left Missing Across
and at End for Survey 2, 57 Items

Number missing	Percent missing	Across all items *f*	Across all items %	At end of survey *f*	At end of survey %
0	0.00	2633	74.4	3459	97.7
1	1.75	574	16.2	33	0.9
2	3.51	175	4.9	4	0.1
3	5.26	52	1.5	1	0.0
4	7.02	19	0.5	1	0.0
5	8.77	16	0.5	4	0.1
>5	more than 10	72	2.0	39	1.1

n= 3541

Survey 3 missing item results are presented in Table 4. Of the 2688, 82.1% answered all of the items, 11.8% had one missing item, and 6.1% had two or more items missing out of the 50 items. Only 2.4% of the respondents had missing items at the end of the survey, with 1.6% having ten percent or more missing at the end of the survey.

Overall for these three surveys, 77.1% of the respondents answered every item. For survey 1, taken by students in grades K-6, 74.4% completed every item; for survey 2, taken by students in grades 5-12, 74.4% completed every item; and for survey 3, taken by teachers, 82.1% completed every item. A comparison of the proportion of surveys with all completed items across these three surveys indicated there was a significant difference, $\chi^2(2, N= 7469) = 58.60, p < .05$. Follow-up pairwise chi-square comparisons, using a Dunn-Bonferroni alpha adjustment, indicated that the teacher group had a higher proportion of complete responses than the K-6 grade group, $\chi^2(1, N= 3928) = 31.46, p < .05$, and the 5-12 grade group, $\chi^2(1, N= 6229) = 52.94, p < .05$.

Table 4
Percent of Items Left Missing Across
and at End for Survey 3, 50 Items

Number missing	Percent missing	Across all items *f*	Across all items %	At end of survey *f*	At end of survey %
0	0	2207	82.1	2623	97.6
1	2.0	317	11.8	18	0.7
2	4.0	75	2.8	2	0.1
3	6.0	30	1.1	1	0.0
4	8.0	8	0.3	1	0.0
>4	10 or more	51	1.9	43	1.6

n= 2688

Comparing the proportion of respondents who left more than 10% of the items blank, 6.2% of the Survey 1 group (grades K-6), 2.0% of the Survey 2 group (grades 5-12), and 1.9% of the survey group (teachers) left more than 10% of the items blank. There was a significant difference in these proportions, $\chi^2(2, N= 7469) = 71.28, p < .05$. Follow-up pairwise chi-square comparisons, using a Dunn-Bonferroni alpha adjustment, indicated that the K-6 grade group had a higher proportion of 10% or more missing items than the 5-12 grade group, $\chi^2(1, N= 4781) = 53.05, p < .05$, and the teacher group, $\chi^2(1, N= 3928) = 50.06, p < .05$.

While only 1.2% of the respondents left more than 10% of the items at the end of the survey blank, the proportion was higher for the two longer surveys. For survey 1, which had 25 items, 0.6% left more than 10% of the items blank at the end of the survey. For survey 2, with 57 items, 1.1% of the respondents left more than 10% of the items at the end of the survey blank and for survey 3, with 50 items, 1.6% of the respondents left 10% or more of the end of survey items blank.

Of the 77 respondents in Survey 1 who had 10% or more items missing, 8 (10.4%) respondents had 10% or more missing at the end of the survey. Of the 72 respondents in Survey 2 who had 10% or more items missing, 39 (54.2%) respondents had 10% or more missing at the end of the survey. Of the 51 respondents in Survey 3 who had 10% or more items missing, 43 (84.3%) respondents had 10% or more missing at the end of the survey. The percentage of 10% or more missing at the end of the survey, as opposed to being missing not all at the end, was significantly different, $\chi^2(2, N= 200) = 71.56, p < .05$. Follow-up pairwise chi-square comparisons, using a Dunn-Bonferroni alpha adjustment, indicated that the teacher group had a higher percentage missing at the end (84.3%) compared with the grade 5-12 group (54.2%), $\chi^2(1, N= 123) = 12.21, p < .05$, and a higher percentage than the K-6 grade group, $\chi^2(1, N= 128) = 69.95, p < .05$. The percentage missing at the end of the survey was higher for the 5-12 grade group compared with the K-6 grade group, $\chi^2(1, N= 149) = 33.02, p < .05$. The difference between the K-6 grade group and the others is likely to be a function of the length of the survey in relation to age. However, the difference between the 5-12 grade group and the teachers cannot be accounted for by differences in the number of items, since actually Survey 2 was longer.

Response Patterns

It may be possible to identify nonattending respondents by examining certain patterns of responses. This was done only with Survey 1 by sorting the subject response mean and standard deviation into the cells which represented low and high standard errors of the item mean and standard deviation. As indicated previously, certain cells may indicate patterns of responses that would be expected from nonattending respondents. Cells 11 and 51 include patterns where the respondent's item mean deviates more than ±1.28 standard errors from the total item mean and the item standard deviation deviates more than -1.28 standard errors from the total item standard deviation. These cells would indicate mono-extreme patterns or situations where the respondent tends to mark only one response at one end of the scale (all SD's for cell 11 and all SA's for cell 51). As indicated in Table 5, these two cells represent 5.4% of the total respondents. As indicated in Table 6, 83.4% of the responses for the 57 respondents in this cell were SA's and 13.9% were A's. Twenty-five of the respondents (2.0%) marked all SA's and eight (0.7%) marked all SA's except for one item. Of the responses in cell 11, 77.3% were SD's and 19.6% were D's. Of the nine respondents in cell 11, six marked all SD's and one marked all D's.

Monotonic patterns at or around the mean of the scale would be represented in cell 31. This cell represents respondents who tend to mark only one response category, that is close to the middle of the scale. For Survey 1, this was 1.6% of the students. Of the entire data set, 32 (2.6%) marked all one response and 11 (0.9%) marked all one response except for one item. Monotonic patterns were indicated for 8.7% of the respondents, with more than three percent marking all or all but one item with the same response.

Table 5
Frequency and Percent of Respondents of z Standard Error Deviations from Item Mean by Item Standard Deviation, Survey 1

z_s deviations→	Very low SD <-1.28		Low SD -1.28 to <-.52		Middle SD -.52 to +.52		High SD >+.52 to +1.28		Very High SD >+1.28		Total by mean category	
z_m deviations↓	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%
Very low mean <-1.28	9	0.7	7	0.6	42	3.4	25	2.0	6	0.5	89	7.3
Low mean -1.28 to <-.52	5	0.4	13	1.1	106	8.7	74	6.1	33	2.7	231	18.9
Middle mean -.52 to +.52	20	1.6	80	6.5	257	21.0	155	12.7	48	3.9	560	45.8
High mean >+.52 to +1.28	15	1.2	50	4.1	131	10.7	32	2.6	1	0.1	229	18.7
Very high mean >+1.28	57	4.7	30	2.5	27	2.2	0	0.0	0	0.0	114	9.3
Total by SD category	106	8.7	180	14.7	563	46.0	286	23.4	88	7.2	1223	

Table 6
Distribution of Item Responses for Selected Item Mean and Standard Deviation z Standard Error Deviation Combinations, Survey 1

Group →	Total	Very low mean Very low SD Cell 11	Very high mean Very low SD Cell 51	Very low Mean Very high SD Cell 15	Low mean Very high SD Cell 25	Middle mean Very high SD Cell 35
n	1223	9	57	6	33	48
Percent missing	1.8	2.2	1.5	3.3	2.2	3.3
Percent SD (1)	18.5	77.3	0.4	58.0	47.2	36.3
Percent D (2)	16.8	19.6	0.8	2.7	4.0	3.8
Percent A (3)	25.8	0.0	13.9	0.7	5.3	4.6
Percent SA (4)	37.2	0.9	83.4	35.3	41.3	52.1
Item *M*	2.83	1.23	3.83	2.14	2.42	2.75
Item *SD*	1.13	0.48	0.43	1.44	1.43	1.42

Monotonic patterns will have item means that are close to the low end (designated as very low mean in Table 7) or high end (designated as very high mean in Table 7) of the scale. Very low means comprised 7.3% of the distribution and 9.3% of the respondents comprised very high means. In the very low mean group, 70.2% of the responses were the lowest two possible responses (SD's or D's) and in the very high mean group, 92.3% of the responses were the highest two categories (A's or SA's). Comparing the percent of respondent item means across the five levels of variation from the mean by grade level (K-2, 3-4, 5-6), there was significant chi-square, $\chi^2(8, N = 1207) = 58.73, p < .05$. Examination of the cell contributions revealed that K-2 students tended to have item means closer to the high point on the scale, students from grades 3 and 4 tended to have means closer to the middle, and students in grades 5 and 6 tended to have means toward the low point on the scale.

BEST COPY AVAILABLE

Table 7
Distribution of Item Responses for Selected Item Mean and
Standard Deviation ± Standard Error Deviations, Survey 1

Group	Total	Very low mean	Very high mean	Very low SD	Very high SD
n	1223	89	114	106	88
Percent Missing	1.8	2.5	1.9	1.9	2.9
Percent SD (1)	18.5	51.8	3.8	7.0	41.7
Percent D (2)	16.8	18.4	2.1	10.1	3.7
Percent A (3)	25.8	13.8	13.3	31.1	4.5
Percent SA (4)	37.2	13.5	79.0	49.9	47.1
Item *M*	2.83	1.89	3.71	3.26	2.59
Item *SD*	1.13	1.10	0.69	0.91	1.44

Patterns with high response variability may also indicate potential nonattending respondents. These are represented by cells 15, 25, and 35. Respondents who mark mostly extreme patterns are referred to as checker-extremes. For cell 15, 58% responded with SD's and 35% responded with SA's; for cell 25, 47% responded with SD's and 41% responded with SA's; and for cell 35, 36% responded with SD's and 52% responded with SA's. For all cells with high item variability (designated as very high SD in Table 7), 42% answered SD's and 47% answered SA's. These respondents represent 7.2% of the total respondent group. While there is not direct evidence that these represent nonattending responses, such responses are suspect when compared with the total respondent group.

Monotonic response patterns will have a standard deviation of item responses which would be low (designated as very low SD in Table 7) and checker-extremes will have a standard deviation of items that would be high (designated as very high SD in Table 7). The relationship of grade level and respondent standard deviation of item responses across the five levels of low to high variation was examined using a chi-square test. There was significant chi-square, $\chi^2(8, N = 1207) = 73.17$, $p < .05$. Examination of the cell contributions revealed that K-2 students tended to have higher spread across the five variation levels; they were more heterogeneous in the spread of item responses. Students in grades 3-6 had standard deviations of items less variable across the five levels; they were more homogeneous compared to the grades K-2 group.

Based on these results (5.4% for mono-extreme, 1.6% for mono-middle, and 7.2% for checker-extreme), more than 14% of the respondents are providing response patterns which are consistent with nonattending patterns. While it is possible that some of these are actually attending to the items, many of them are likely to be not attending. Such patterns will have different effects on commonly used questionnaire statistics, particularly on Cronbach's alpha. Mono-extremes increase alpha, checker-extremes reduce alpha, and mono-middles have a negligible effect on alpha (Barnette, 1996).

Item Reversals

Surveys 2 and 3 had items which were reverse-worded. It would be assumed that the distribution of nonreversed items should be about the same as the distribution of reverse-worded items after the score scale was reversed. This analysis was based on comparing these distributions using $df = 1$, chi-square tests with the strongly disagree and disagree responses collapsed into one response category and the strongly agree and agree responses collapsed into the other response category. Thus, the chi-square was conducted on a 2X2 matrix: reversed and nonreversed items by agree and disagree. Observed chi-square values were computed and probabilities determined for each respondent.

Table 8 presents the results of this analysis. Survey 2 was completed by students, and Survey 3 was taken by teachers. For Survey 2, which had 14 reverse-worded items out of 57, 68.7% did not have a significant difference at $p < .05$. However, there were significant differences at $p < .05$ for 31.3% of the respondents. At the $p < .01$ level, 17.7% of the respondents were significantly different, and at $p < .001$, 9.7% had significant differences. Comparison of the percentages having significant differences at the .05 level among the three student grade levels: elementary (30.7%), middle (32.5%), and secondary (31.2%), failed to find a significant difference, $\chi^2(2, N = 3520) = .64$, $p > .05$. Therefore, the percentage of Survey 2 respondents (students in grades 5-12) providing significantly different direct and reverse-worded item distributions is not related to student grade level.

Table 8
Differences Between Direct and Reverse-Worded Item Distributions for Surveys 2 and 3. Percent of Chi-Square Probabilities

	n	%, $p >= .05$	%, $p < .05$	%, $p < .01$	%, $p < .001$
Survey 2	3520	68.7	31.3	17.7	9.7
Survey 3	2658	74.2	25.8	10.3	1.6

Survey 3, which had 19 reverse-worded items out of 50, had no significant differences at $p < .05$ for 74.2% of the respondents. There were 25.8% with significant differences at $p < .05$, 10.3% had significant differences at $p < .01$, and 1.6% had significant differences at $p < .001$. Comparison of the percentages with significant differences at the .05 level among the three school levels: elementary (28.0%), middle (24.9%), and secondary (23.9%), failed to find a significant difference, $\chi^2(2, N = 2568) = 5.11, p > .05$. Therefore, the percentage of Survey 3 respondents (teachers) providing significantly different direct and reverse-worded item distributions is not related to student grade level.

A substantial proportion of respondents provided different levels of disagree-agreement on the direct worded items compared with the reverse-worded items. This is more the case with the student group compared with the teacher group. Some of this may be due to providing monotonic patterns of responses without paying attention to the reverse wording, clearly nonattending behavior. However, another possible reason for some of these differences may be the tendency to be more willing to respond with a "strongly agree" response compared with a "strongly disagree" response. Reverse-worded items would convert a "strongly agree" score point to a "strongly disagree" score point. When asked directly, the respondent may only be willing to mark a "disagree" rather than a "strongly disagree."

Conclusions and Discussion

Based on these results, the proportion of missing items does not seem to be a major problem. Most of the respondents (77.1%) answer all of the items. Only about 2.7% leave more than 10% of the items blank. Teachers are more likely to answer all of the items than are students. Elementary students are more likely to leave 10% or more of the items blank. A very small percentage (1.2%) of the respondents leave ten percent or more missing at the end of the survey. Teachers are more likely to leave 10% or more of the items blank at the end of the survey as compared with students. If teachers do not respond to all the items, they are more likely than students to leave those items blank at the end of the survey.

It would be important to replace missing items in order to compute Cronbach's alpha based on a more complete data set. These three data sets lost almost 23% of the subjects in computations of Cronbach's alpha due to missing items. Also, if total scores or subscale scores are needed, based on summing item responses, the missing items would have to be accounted for. However, these results do not seem to indicate a high occurrence of nonattending behaviors associated with leaving items blank for these three survey data sets.

While there is often a need to account for missing responses for purposes of determining reliability with complete data sets or to determine subscale or total score for inferential data analysis purposes, typical methods are likely to be sufficient for dealing with this problem. However, as Witta (1994) has pointed out, different replacement methods may be less useful when data are not missing randomly. Clearly, items left blank at the end of a survey would not be missing randomly.

There was evidence of response patterns which may be associated with nonattending behaviors. The most obvious patterns of monotonic responses and checker-extreme patterns were present in Survey 1. More than five percent of the respondents had all or a very high proportion of responses at all ones or all fours, evidence of possible monotonic behaviors. For more than seven percent of the respondents there was an almost even balance of ones (42%) and fours (47%) with only eight percent of the responses at two or three, a checker-extreme pattern. It seems likely that, to at least some degree, many of these could be nonattending respondents. Such response patterns, if they are from nonattending respondents, introduce error into the data set, affecting reliability, standard error of measurement, and statistics that may be used for inferential comparisons such as the survey mean and variance. Some of these patterns may have affects that cancel each other out, but as Barnette (1996) has pointed out, the mono-extreme pattern, even at low levels of occurrence, has a strong influence on Cronbach's alpha. Such a pattern spuriously inflates alpha, leading to believing the survey is more reliable than it actually is.

Surveys 2 and 3 had reverse-worded items. Comparing the distributions of these two types of items, after reverse scoring the reverse-worded items, a high percentage of respondents (31.4% for Survey 2 and 25.7% for Survey 3) had significant differences between reverse and non-reverse-worded item distributions at

$p < .05$. Thus, by chance, we would expect about five percent to be significantly different and the result here was much higher. Based on these results, it seems that there may be more than 25% of the respondents who may not be attending to reverse wording of survey items. Grade level of student and teaching level are not related to the tendency to not attend to reverse-worded items.

Second only to achievement testing, the use of attitude surveys for educational data collection is frequent and pervasive. They are used for program evaluation, individual attitude assessment, parent satisfaction, faculty or staff evaluation, and for collecting research data. Educational decisions need to be based on accurate, reliable, and valid data collection using the best instruments available, completed by respondents who provide thoughtful and attentive responses.

Implications for Further Research

Valid and reliable data from attitude surveys are highly desirable. Any respondent behavior that introduces error needs to be identified and removed, or at least the effects minimized. The three types of potential nonattending behaviors discussed here pose different problems of identification and treatment. Relative to missing items, we need to identify why they occur, what differences there are between missing items throughout the survey compared with those occurring at the end of the survey, and the best methods to use to deal with these different situations. One issue which needs to be addressed when items are missing at the end of the survey is the internal consistency of items up to the point of discontinuation. If there is reasonably high internal consistency, then item replacement is a reasonable alternative. A related issue is which replacement methods are better to use to replace items missing at the end of the survey. However, if there is low or no internal consistency, then the respondent should be deleted.

Response patterns such as those identified in this research and others need to be examined relative to how frequently they occur, how they may be identified, what effects they have, and how to deal with them in real data sets. Studies need to be conducted to determine if these observed patterns and item reversals are really related to nonattending behaviors. If respondents are not attending to reverse-worded items, we need to determine characteristics of respondents who do not or can not deal with negatively worded items.

If respondents are not attending, are these behaviors related to length of survey, captive vs. non-captive administration settings, anonymity of respondent, use of optical scanning devices vs. marking directly on the form, and different nonattending patterns on early vs. later items? Also, what are the effects of nonattending response patterns on commonly used survey statistics: internal consistency reliability, survey means and variances, and effect sizes? This is an area of research which has not received very much attention, yet when we consider how often we use these instruments, it would seem critical that we determine the extent to which such behaviors exist, how to identify them, determine their effects, and consequences of keeping them in the data base.

References

Barnette, J. J. (1996, April). *Simulated nonattending respondent effects on internal consistency of self-administered questionnaires.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Bradburn, N. M., & Sudman, S. (1991). The current status of questionnaire design. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 29-40). New York: John Wiley & Sons.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Goldsmith, R. E. (1986). Personality and uninformed response error. *Journal of Social Psychology*, 126(1), 37-45.

Goldsmith, R. E. (1988). Spurious response error in a new product survey. *Journal of Business Research, 17*, 271-281.

Goldsmith, R. E. (1989). Reducing spurious response in a field survey. *Journal of Social Psychology, 129*(2), 201-212.

Groves, R. M. (1991). Measurement error across the disciplines. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 1-25). New York: John Wiley & Sons.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs: Prentice-Hall.

Huberty, C. J., & Julian, M. W. (1995). An ad hoc analysis strategy with missing data. *Journal of Experimental Education, 63*(4), 333-342.

Kaiser, J. (1990, August). *The robustness and substitution by mean methods in handling missing values.* A paper presented at the annual Islamic conference on statistical sciences, Johor Bahru, Malasia. (ERIC Document Reproduction Service No. ED 325 492)

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys.* New York: John Wiley & Sons.

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*(1), 37-49.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Pilotte, W. J., & Gamble, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*, 603-610.

Wentland, E. J., & Smith, K. W. (1993). *Survey responses: An evaluation of their validity*. San Diego: Academic Press.

Witta, E. L. (1994, November). *Are values missing randomly in survey research*. A paper presented at the annual meeting of the Mid-South Educational Research Association, Nashville, TN. (ERIC Document Reproduction Service No. ED 389 727)

Witta, E. L., & Kaiser, J. (1991, November). *Four methods of handling missing data with the 1984 General Social Survey*. A paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, KY. (ERIC Document Reproduction Service No. ED 339 755)

Copyright 1996 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1996, Vol. 3, No. 2, 61-74

Influences on and Limitations of Classical Test Theory Reliability Estimates

Margery E. Arnold
Texas A&M University

The present paper explains how different factors affect classical reliability estimates such as test-retest, inter-rater, internal consistency, and equivalent forms coefficients. Furthermore, the limits of classical test theory are demonstrated, and it is recommended that researchers, teachers and psychologists instead learn and use generalizability-theory estimates of reliability. Concrete examples and detailed explanations make this discussion accessible even to those who are uninitiated in either classical test theory or generalizability theory.

The reliability of test scores concern teachers, psychologists and researchers who want to know that the scores on the tests which they administer are consistent and generalizable. Unfortunately, many training programs in the disciplines of education and psychology still emphasize classical methods of deriving reliability coefficients (such as test-retest reliability, internal consistency reliability), when more advanced methods are now available. In addition, typical ways of teaching about selecting and using standardized tests may unwittingly teach students that instruments or tests can possess a quality called "reliability." Such teaching practices may lead to misuse of tests and misinterpretation of test scores. Because test scores can have such immense impact upon the lives of students or clients, it is very important that the misuse and misinterpretation of tests be avoided.

The purpose of the present paper, therefore, is to teach in an easy-to-understand manner, with illustrations that are well-chosen and plentiful, the following concepts: the limitations of classical test theory reliability estimates, a more advanced method of estimating reliability which is known as generalizability theory, and the dangers of believing that reliability is a quality of a test when, in fact, reliability is a quality of *scores* on the test, not a quality of the test itself. The present paper is organized in the following manner. First, the present paper reviews classical test theory reliability coefficients and their corresponding sources of measurement error (inconsistencies in occasions, forms, raters and items). In addition, the first section explains how to compute the reliability coefficient that corresponds to each source of error. The paper then explains different factors that can affect classical test theory reliability estimates, illustrating that reliability is a quality of *scores* on tests, and *not* a quality of tests or instruments. Then, the paper relates two reasons that teachers, psychologists and researchers should calculate reliability estimates on each set of test scores that they obtain. Finally, two problems associated with classical test theory reliability estimates are outlined and a suggestion is made for the use of generalizability theory reliability estimates instead of classical test theory reliability estimates.

Definitions of Reliability According to Classical Test Theory

To explain the factors that limit reliability coefficients, it is helpful to give a brief review of the concept of reliability. Reliability expresses the relationship between observed scores and true scores. A concrete example using the spelling test scores of a second grade class clarifies these constructs. "True score" refers to each class member's true ability in the domain of second grade spelling. "Observed score" refers to each class member's actual score on the spelling test. Thus, reliability concerns the relationship between what the children actually know about spelling (true score) and what they made on their spelling test (observed score). This relationship between true score and observed score can be conceptually explained as a mathematical model, a statistic, and an illustration, all of which are demonstrated in the following discussion.

The relationship between true score and observed score as expressed in a mathematical model was explained by Charles Spearman (1907, 1913, cited in Crocker & Algina, 1986) in what has become known as the true score model or classical test theory. According

---

Margery Arnold is a doctoral candidate in Counseling Psychology in the Department of Educational Psychology at Texas A & M University where she is supported by a Regents Fellowship. She wishes to acknowledge and thank her advisor, Dr. Bruce Thompson, for his encouragement and guidance on this article. Correspondence regarding this article should be sent to Margery E. Arnold, Department of Educational Psychology, Texas A & M University, Mail Stop 4225, College Station, TX 77843-4225, E-mail: mea8714@acs.tamu.edu.

to Crocker and Algina (1986), "the essence of Spearman's model was that any observed test score could be envisioned as the composite of two hypothetical components--a true score and a random error component" (p. 107). Thus, the equation is Observed score = True score + Random error. In other words, observed scores are composed of true scores (which, by definition are reliable) and an error component that is not reliable. From the true score model one can construct another equation for reliability as expressed in a ratio of true score variance to observed score variance. Therefore, reliability is "the proportion of observed score variance that may be attributed to variation in the examinees' true scores" (Crocker & Algina, 1986, p. 116).

It is important to note that the error term in this model can have a positive effect or a negative effect on observed scores. For example, if a student does not know the answer to a question, but guesses correctly, his or her observed score will be higher than his or her true score. In this case the measurement error that was introduced by the guess has a *positive* effect on the observed score. Alternatively, another student might know the correct answer and mismark the answer, so that she or he answers incorrectly. In this case, the true score exceeds the observed score, and the error that was introduced by mismarking had a *negative* effect on the student's observed score.

As previously noted, reliability can also be expressed as a statistic such as coefficient alpha or as the Pearson product moment correlation, *r*. Alpha will be explained later, so this demonstration focuses on *r*. Expressed as a correlation, reliability is the correlation between the true scores (actual ability in the domain of second grade spelling) and the observed test scores (the class's scores on the spelling test). [The correlation between true scores and observed scores is called the reliability index. The reliability coefficient is the reliability index squared (Crocker & Algina, 1986, pp. 115-116).] This relationship between true scores and observed scores can also be illustrated as the graph found in Figure 1, or the diagram found in Figure 2.

Methods of Estimating Measurement Error Using the True Score Model

Crocker and Algina (1986) defined true score "as the average of the observed scores obtained over an infinite number of repeated testings with the same test" (p. 109). Unfortunately it is impossible (and impractical) to calculate true scores in this manner. True scores cannot be exactly calculated. They can only be estimated. True scores are estimated using what the researcher can obtain--observed scores, measurement error estimates (i.e., reliability coefficients) and the true score model.

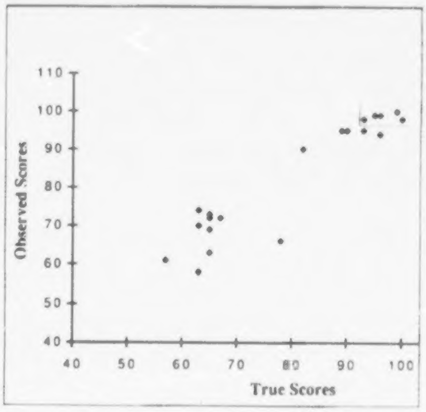


Figure 1.

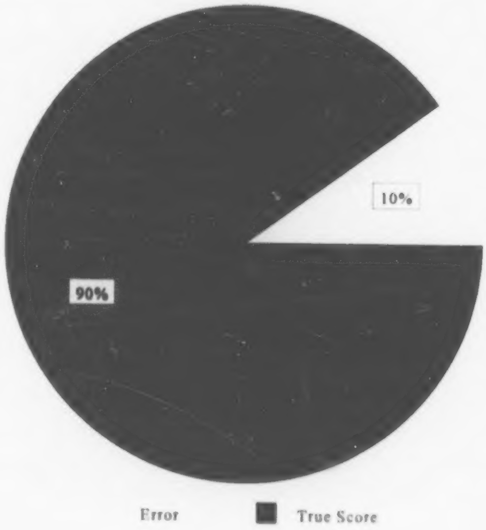


Figure 2.

The true score is predicted by estimating the amount of measurement error that occurred in the administration of a test and then adjusting the observed scores using that estimation of error. If one knows the measurement error, then one can estimate the extent to which the measurement error has caused the observed scores to deviate from the true scores. It follows, then, that the estimation of measurement error is the key to finding true scores. Crocker and Algina (1986) defined error in the classical true score model as "an error of measurement. . . the discrepancy between an examinee's observed test score and his or her true score" (p. 110). [Note that this type of error is *measurement error* as opposed to *sampling error* or *model specification error*. Sampling error is the difference between the statistic one obtained by measuring a sample and the statistic that one would have obtained had one sampled the entire population. Model specification error is the variance in the observed dependent variable scores that is not explained due to the specification of the wrong predictive model. Said differently, model specification error is the variance in the observed dependent variable(s) that is not explained by the independent variable(s).]

In classical test theory, there are four sources of measurement error that are often estimated: (a) inconsistencies in occasions, (b) inconsistencies in forms, (c) inconsistencies among raters, and (d) inconsistencies in sampling the content domain. What follows is a discussion for each source of error explaining how each of these inconsistencies is a source of measurement error and how to compute the reliability coefficient that corresponds to that source of error. Before moving directly to those explanations, however, it is important to note Gronlund's (1976) warning regarding reliability and particular sources of error:

> An estimate of reliability always refers to a particular type of consistency [e.g., consistency across occasions or across forms or across raters or across items sampled]. . . . It is possible for test scores to be consistent in one of these respects and not in another. The appropriate type of consistency in a particular case is dictated by the use to be made of the results. . . . The reliability coefficient resulting from each method [of calculating reliability] must be interpreted in terms of the type of consistency being investigated. (pp. 106-108)

Consistency Across Time

With this warning having been heeded, an explanation of different sources of error and how to calculate reliability coefficients for each of those sources is considered. A researcher may be concerned with how stable his or her observed test scores will be over time. In other words, if the test were administered to the same group of people on a future occasion, how different will the test scores obtained on the second occasion be from the scores observed on the first occasion? The difference between the scores on these two occasions is a source of measurement error--measurement error due to occasions. Once this source of error has been measured (by testing the same group of persons with the same test on two different occasions) one can compute a reliability coefficient called the "stability coefficient" or "test-retest reliability." The stability coefficient is calculated by computing the Pearson product moment correlation between the scores on the two different occasions (Crocker & Algina, 1986).

Consistency Across Forms

A second source of measurement error comes from inconsistencies in test forms. A teacher who would like to deter cheating on an exam may give two different forms of the same test. To understand how giving two different forms of the test might have introduced measurement error into the observed scores, the teacher may compute a reliability coefficient called the "equivalence coefficient." The equivalence coefficient is computed by calculating the Pearson product moment correlation between the scores on the two forms. Note that to compute this coefficient it is necessary to give both forms to at least a portion of the persons taking the test. One form is given and then the second form is administered within a short period of time. Usually the order of the forms is counterbalanced, so that order of test form will not affect scores or reliability estimates.

Consistency Across Raters

A third source of measurement error arises from inconsistencies between raters. (This type of error, of course, only occurs when one uses raters and will not occur during the administration of an objective exam with accurate machine scoring.) For example, one rater may have a slightly different method of assigning scores than another rater. To estimate the extent to which this type of measurement error has differentiated the observed scores from the true scores, one may calculate the inter-rater percent agreement, or some similar coefficient.

Consistency Across Items

A fourth major source of measurement error in classical test theory arises from inconsistencies in sampling the content domain or from inconsistencies in items.

When a teacher (or psychologist) gives a test, he or she cannot possibly ask all of the questions in the domain of the content area being tested. Therefore, he or she must select possible items from the larger content domain. The teacher or psychologist hopes that he or she selected the correct items such that scores on the test he or she created can generalize to the domain of questions that might have been asked (Crocker & Algina, 1986). One may calculate the "internal consistency reliability coefficient" to estimate the extent to which this type of measurement error has caused the observed scores to deviate from the true scores. There are several ways to compute the internal consistency coefficient; all are based on the correlation between separately scored parts of the test (Crocker & Algina, 1986). "If examinees' performance is consistent across subsets of items within a test, the examiner can have some confidence that this performance would generalize to other possible items in the content domain" (Crocker & Algina, 1986, p. 135). Three common ways to calculate internal consistency reliability are the split-half coefficient, Kuder Richardson 20 (KR 20) and Cronbach's (1951) alpha. Split-half coefficients are the Pearson product-moment correlations between scores on two halves of the same test. KR 20 and Cronbach's alpha are computed with similar formulas. Notice that the formulas are identical with one exception--how they compute the sum of the item variances. This difference arises because KR 20 is used only with dichotomously scored data. The use of only dichotomously scored data provides for a simpler formula for item variance.

$$\text{KR 20} = [k/(k-1)][1-(\Sigma(pq)/\sigma_x^2)]$$
$$\text{Cronbach's Alpha} = [k/(k-1)][1-(\Sigma\sigma_i^2/\sigma_x^2)]$$

k = number of items
p = proportion of persons answering the item correctly
q = proportion of persons answering the item incorrectly
$\Sigma\sigma_i^2$ = sum of the item score variances
σ_x^2 = test score variance

KR 20 is, in effect, a *special* case of Cronbach's alpha. KR 20 calculates internal consistency reliability for the specific situation of all test items being scored 0 or 1 ("dichotomously"); p times q is merely an algebraically simpler way to compute the variance of dichotomous item scores. Conversely, Cronbach's alpha is the *generic* case of KR 20. Cronbach's alpha calculates internal consistency reliability coefficients for multipoint items (such as multiple choice items) as well as dichotomously scored items.

Factors Affecting the Magnitude of Reliability Coefficients

Several factors influence the magnitude of reliability coefficients. Among those elements are time limits placed on the test, the spread or variability of the scores, the length of the test and the difficulty of the items, as well as the examinees themselves. The following section uses a spreadsheet program created from the formula for KR 20 to demonstrate how time limits, test score variability, test length and item difficulty affect reliability coefficients. Finally, another example is given to explain how the examinees may affect coefficient alpha.

Time Limits

To understand how the amount of time allowed for a person to take a test can affect coefficient alpha, it is important to recall that the classical true score model assumes that measurement error is random, not systematic (Observed Score = True Score [*systematic*] + Error [*random*]). The speed at which an examinee can complete a test is attributible to individual differences. Therefore, speed is an ability that would fall under the systematic (true score) part of the true score model rather than under the random (measurement error) part of the the true score model.

Tables 1 and 2 demonstrate the effects of speed on a seven-item, reading comprehension test taken by 10 persons. {Note. The grid represents persons' scores (0=incorrect and 1=correct) on the seven items. The final column lists each individual's total score. Across the bottom of the grid one can find p or the difficulty for each item and v the variance of each item. Below the grid one can find the elements of the KR-20 formula for alpha [*k/k-1*][*1-* (the sum of the item variances/ total test score variance)].} In Table 1, the examinees were given as much time as they wanted to complete the test. In Table 2, the test was timed and four members of the class--Todd, Nancy, Lu and Tammi--did not do as well as they did when they had all the time that they needed. (The items that they missed are in bold.)

One could reasonably argue that the first, untimed test scores illustrate the abilities of the class on reading comprehension better than do the timed test scores. Nonetheless, the reliability of the scores on the timed test (α = .74) is greater than the reliability of scores on the untimed test (α = .20). This occurs because the timed test measures two abilities--reading comprehension *and* speed--as opposed to the untimed test which only measures reading comprehension. This measuring of two abilities provides for a greater spread in the total test scores. (Notice that the range on the timed test is 7-1=6, while the range on the untimed test is 7-4=3.)

Table 1
Ten Students' Scores on a 7-item, Nonspeeded Test with Coefficient Alpha Calculations

	1	2	3	4	5	6	7	Score	
Todd	1	1	0	1	1	1	0	5	
Lu	0	1	1	1	1	1	1	6	
Nancy	1	1	1	1	1	1	1	7	
Tammi	1	1	0	1	0	1	0	4	
Karen	1	1	1	1	1	0	1	6	
Avery	0	1	1	1	1	1	1	6	
Art	1	0	1	0	1	0	1	4	
Cris	1	1	0	1	1	1	1	6	
Kris	1	1	1	1	1	1	1	7	
Brad	1	1	1	1	1	1	1	7	
P	0.8	0.9	0.7	0.9	0.9	0.8	0.8	5.8	*M*
V	0.16	0.09	0.21	0.09	0.09	0.16	0.16	1.16	Total test variance
								1.17	K/K-1
		α= 0.20						0.96	Sum of item variances
								0.83	sum of item variances/total test variance
								0.17	1- (sum of item variances/total test variance)

Table 2
Ten Students'scores on a 7-item, Speeded Test with Coefficient Alpha Calculations

	1	2	3	4	5	6	7	Score	
Todd	1	1	0	**0**	**0**	**0**	0	**2**	
Lu	0	1	1	1	1	1	**0**	5	
Nancy	1	**0**	**0**	**0**	**0**	**0**	**0**	1	
Tammi	1	1	0	1	0	**0**	0	**3**	
Karen	1	1	1	1	1	0	1	6	
Avery	0		1	1	1	1	1	6	
Art	1	0	1	0	1	0	1	4	
Cris	1	1	0	1	1	1	1	6	
Kris	1	1	1	1	1	1	1	7	
Brad	1	1	1	1	1	1	1	7	
P	0.8	0.8	0.6	0.7	0.7	0.5	0.6	4.7	*M*
V	0.16	0.16	0.24	0.21	0.21	0.25	0.24	4.01	Total test variance
								1.17	K/K-1
		α= 0.74						1.47	Sum of item variances
								0.37	sum of item variances/total test variance
									1- (sum of item variances/total test variance)

Variability of Test Scores

The spread, or variance, in total test scores is the element of the KR-20 equation that *most greatly affects* the magnitude of coefficient alpha (Reinhardt, 1991). The spreadsheet program used in the timed-test example can also be used to illustrate how total test score variance affects coefficient alpha. Table 3 illustrates that when there is very little test score variance, coefficient alpha is at an absolute low (α = -21). **Note that coefficient alpha can be negative**. (This occurs when the sum of the item variances is greater than the total test score variance.) Also note that if there is no variability in total

BEST COPY AVAILABLE 455

test scores, then it is impossible to compute coefficient alpha. (It is impossible to divide by zero.) Therefore, in Table 3, the test score variance has been reduced to the lowest possible level without becoming zero.

Table 3 was transformed (changed values are in bold in Table 4) to create maximum test score variance. The item responses were changed so that half of the students answered all of the items correctly, while the other half of the students answered all of the items incorrectly. Arranging the test scores this way creates maximum deviation from the mean test score. (Recall that the formula for variance has as its numerator the sum of the squared deviations from the mean.) While these test scores are probably not test scores desired by any teacher, they are the test scores that will produce maximum total test score variance, and thus, maximum coefficient alpha.

Table 3
Minimal Test Score Variance Leads to Minimal Coefficient Alpha

	1	2	3	4	5	6	7	Score	
Buzz	0	0	0	0	0	0	0	0	
Meg	0	0	0	0	0	0	0	0	
Skip	1	1	1	1	0	0	0	4	
Jan	1	1	1	0	0	0	1	4	
Mark	1	1	0	0	0	1	1	4	
Joy	1	0	0	0	1	1	1	4	
Max	0	0	0	1	1	1	1	4	
Lucy	0	0	1	1	1	1	0	4	
Alex	0	1	1	1	1	0	0	4	
Gina	1	1	0	0	1	1	0	4	
P	0.5	0.6	0.6	0.6	0.6	0.6	0.4	3.9	*M*
V	0.25	0.24	0.24	0.24	0.24	0.24	0.24	0.09	Total test variance
								1.17	K/K-1
	α= -21							1.69	Sum of item variances
								18.8	sum of item variances/total test variance
								-18	1- (sum of item variances/total test variance)

Table 4
Maximal Test Score Variance Leads to Maximal Coefficient Alpha

	1	2	3	4	5	6	7	Score	
Buzz	0	0	0	0	0	0	0	0	
Meg	0	0	0	0	0	0	0	0	
Skip	**0**	**0**	**0**	**0**	0	0	0	0	
Jan	**0**	**0**	**0**	0	0	0	**0**	0	
Mark	**0**	**0**	0	0	0	**0**	**0**	0	
Joy	1	**1**	**1**	**1**	1	1	1	7	
Max	**1**	**1**	**1**	1	1	1	1	7	
Lucy	**1**	**1**	1	1	1	1	**1**	7	
Alex	**1**	1	1	1	1	**1**	**1**	7	
Gina	1	1	**1**	**1**	1	1	**1**	7	
P	0.5	0.5	0.5	0.5	0.5	0.5	0.5	3.5	*M*
V	0.25	0.25	0.25	0.25	0.25	0.25	0.25	12.3	Total test variance
								1.17	K/K-1
	α= 1							1.75	Sum of item variances
								0.14	sum of item variances/total test variance
								0.86	1- (sum of item variances/total test variance)

The previous description mathematically explains how test score variance can increase reliability. Gronlund (1976) offered a conceptual explanation:

> Since the larger reliability coefficients result when individuals tend to stay in the same relative position in a group, from one testing to another, it naturally follows that anything which reduces the possibility of shifting positions in the group also contributes to larger reliability coefficients. In this case greater differences between the scores of individuals reduce the possibility of shifting positions. (p. 118)

Length of the Test

A related concept concerns the effects of length of test on reliability coefficients. Longer tests, generally speaking, are likely to create more test score variance and thus increase reliability coefficients. It is possible, however, for one test to be longer than a second test and still yield scores with the exact same or lower reliability estimates than the shorter test. "There is one important reservation in evaluating the influence of test length on the reliability of the scores, . . . [this rule] . . . assume[s] that the test will be lengthened by adding test items of the same quality as those already in the test" (Gronlund, 1976, p. 118).

In other words, if the items added to the test are worse than the items already on the test, the longer test may actually yield lower reliability coefficients than the shorter test. Tables 5 through 8 demonstrate how adding items to a test may make reliability better, worse or the same, depending upon the quality of the added items.

Table 5 lists the scores of 10 persons on a seven item test (**α**= .84). Two items were added to this test and scores on the new items can be seen in Table 6. Notice that the two items that are added do not change the rankings of the examinees. Everyone answered the two added items incorrectly. Notice also that the coefficient alpha exactly equals the alpha of the test without the added items (**α**= .84).

A third example is given in Table 7. Notice that the added items increased the spread of the test scores; the range is now 9 instead of 7. This increase in total test score variance increased coefficient alpha (**α**= .89).

However, in Table 8 one can see how adding items of lesser quality than the original items can actually *decrease* coefficient alpha (**α**= .69). In this example, the added items were of lesser quality than the original items because persons who had scored low (Skip and Jan) on the original test answered the items correctly, while persons who scored high (Alex and Gina) on the original test answered the added items incorrectly. Furthermore, the added items decreased the variance of the total test scores. (Notice that the range on the original test was 7, while the range on the new test is 6.)

Table 5
Ten Persons' Scores on a Seven Item Dichotomously Scored Test

	1	2	3	4	5	6	7	Score	
Buzz	0	0	0	0	0	1	1	2	
Meg	0	0	0	1	0	0	1	2	
Skip	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	1	1	
Mark	0	0	0	0	0	1	1	2	
Joy	0	0	0	0	1	1	1	3	
Max	0	0	0	1	1	1	1	4	
Lucy	0	0	1	1	1	1	1	5	
Alex	0	1	1	1	1	1	1	6	
Gina	1	1	1	1	1	1	1	7	
P	0.1	0.2	0.3	0.5	0.5	0.7	0.9	3.2	*M*
V	0.09	0.16	0.21	0.25	0.25	0.21	0.09	4.56	Total test variance
								1.17	K/K-1
	α= 0.84							1.75	Sum of item variances
								0.14	sum of item variances/total test variance
								0.86	1- (sum of item variances/total test variance)

Table 6
Scores from Tests in Table 7 with Two Added Items That Everyone Answers Incorrectly

	1	2	3	4	5	6	7	8	9	Score	
Buzz	0	0	0	0	0	1	1	0	0	2	
Meg	0	0	0	0	0	0	1	0	0	1	
Skip	0	0	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	1	0	0	1	
Mark	0	0	0	0	0	1	1	0	0	2	
Joy	0	0	0	0	1	1	1	0	0	3	
Max	0	0	0	1	1	1	1	0	0	4	
Lucy	0	0	1	1	1	1	1	0	0	5	
Alex	0	1	1	1	1	1	1	0	0	6	
Gina	1	1	1	1	1	1	1	0	0	7	
P	0.1	0.2	0.3	0.4	0.5	0.7	0.9	0	0	3.1	*M*
V	0.09	0.16	0.21	0.24	0.25	0.21	0.09	0	0	4.89	Total test variance
										1.13	K/K-1
		α=0.84								1.25	Sum of item variances
										0.26	Sum of item variances/total t
										0.74	1- (sum of item variances /t)

Table 7
Scores from Test in Table 7 with Two Added Items That Add Score Variability

	1	2	3	4	5	6	7	8	9	Score	
Buzz	0	0	0	0	0	1	1	0	0	2	
Meg	0	0	0	0	0	0	1	0	0	1	
Skip	0	0	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	1	0	0	1	
Mark	0	0	0	0	0	1	1	0	0	2	
Joy	0	0	0	0	1	1	1	0	0	3	
Max	0	0	0	1	1	1	1	0	0	4	
Lucy	0	0	1	1	1	1	1	0	0	5	
Alex	0	1	1	1	1	1	1	0	0	6	
Gina	1	1	1	1	1	1	1	**1**	**1**	9	
P	0.1	0.2	0.3	0.4	0.5	0.7	0.9	0.1	0.1	3.3	*M*
V	0.09	0.16	0.21	0.24	0.25	0.21	0.09	0.09	0.09	6.81	Total test variance
										1.13	K/K-1
			α=0.89							1.43	Sum of item variances
										0.21	Sum of item variances/total t

Table 8
Scores from Tests in Table 7 with Two Added Items That Decrease Score Variability

	1	2	3	4	5	6	7	8	9	Score	
Buzz	0	0	0	0	0	1	1	0	0	2	
Meg	0	0	0	0	0	0	1	0	0	1	
Skip	0	0	0	0	0	0	0	**1**	**1**	2	
Jan	0	0	0	0	0	0	1	**1**	**1**	3	
Mark	0	0	0	0	0	1		0	0	2	
Joy	0	0	0	0	1	1	1	0	0	3	
Max	0	0	0	1	1	1	1	0	0	4	
Lucy	0	0	1	1	1	1	1	0	0	5	
Alex	0	1	1	1	1	1	1	0	0	6	
Gina	1	1	1	1	1	1	1	1	1	7	
P	0.1	0.2	0.3	0.4	0.5	0.7	0.9	0.2	0.2	3.5	*M*
V	0.09	0.16	0.21	0.24	0.25	0.21	0.09	0.16	0.16	3.45	Total test variance
										1.13	K/K-1
		α=**0.61**								1.43	Sum of item variances
										0.21	Sum of item variances/total t
										0.79	1- (sum of item variances /t)

Item Difficulty

The item difficulty affects reliability in much the same way as test length does--by increasing or decreasing total test score variance. If all of the items on a test are rather difficult for all of the examinees, then the test score variance will be small and the reliability coefficient will be low. (The range of scores will be restricted, with everyone scoring near 0% correct.) The same phenomenon occurs if all of the examinees answer almost all of the items correctly. Reliability will be low because test score variance is low. (The range of scores will be restricted, with everyone scoring near 100% correct.) If, however, the test is of a medium difficulty for the examinees, the scores will have a greater range, and reliability will be increased. Gronlund (1976) explained that to maximize reliability one should design a test on which

> the average score is 50 percent correct and that the scores range from near zero to near perfect. . . . We can estimate the ideal average difficulty for a selection-type test by taking the point midway between the expected chance score and the maximum possible score. Thus for a 100 item true-false test the ideal average difficulty would be 75 (midway between 50 and 100), and for a 100 item five-choice multiple choice test the ideal average difficulty would be 60 (midway between 20 and 100). (p. 121)

Examinees

The examinees also affect the magnitude of reliability coefficients. Recall that the reliability coefficient has been illustrated as the correlation statistic, r. Hinkle, Wiersma, and Jurs (1994) have described how r is affected by homogeneity of the examinee group. When a group of examinees is very homogeneous all of the members tend to score similarly to one another. Because the range of scores is very small, the standard deviation of the scores is very small. After reflecting on the formula for the correlation coefficient, r = Covariance of X & Y /[(standard deviation of X)(Standard deviation of Y)], one notices that,

> If a group is sufficiently homogeneous on either or both variables, the variance (and hence the standard deviation) tends toward zero. . . . When this happens, we are dividing by zero, and the formula becomes meaningless. In essence, the variable has been reduced to a constant. As a group under study becomes increasingly homogeneous, the correlation coefficient decreases. (Hinkle, Wiersma & Jurs, 1994, pp. 115-116)

Because reliability is a correlation coefficient, it is affected by the homogeneity of the group to whom the test is given. As the group being studied becomes increasingly homogeneous, the reliability coefficient decreases. Thus reliability is affected not only by the properties of the items on the test, but also by the persons taking the test.

An illustration using the spelling test example makes this more clear. Figure 1 displays the relationship between the entire second grade's *true* spelling scores and their *observed* spelling test scores. The reliability of the spelling scores is calculated in Table 9 to be .89, which is considered to be reliable.

Table 9
Reliability Calculated as the Correlation Between True Scores and Observed Scores

2nd Graders	True Score	Observed Score
Jane	57	61
Julio	93	98
Max	63	70
Maria	100	98
Jason	63	58
Leticia	99	100
Margaret	65	72
Anna	96	99
Michael	65	63
David	95	99
Emily	65	69
Sara	96	94
Cathy	78	66
Arnoldo	89	95
Ramiro	67	72
Wayne	93	95
Mitchell	63	74
Matthew	90	95
April	65	73
Amy	82	90
Joan	65	75
Craig	83	90
Linda	75	65
Ruth	85	82
Todd	72	83
Andrew	72	84
Fred	79	88
Juan	82	80
Virginia	78	85
Sum	2275	2373
Mean	78.45	81.83
Reliability of the Entire Class		r_{xx} = .89
Reliability of the top 5 persons in the class (bolded names)		r_{xx} = .27

A second reliability coefficient has also been calculated in Table 9. This second coefficient is the reliability coefficient for the scores of the top five students in the second grade. The top five students are likely to do homogeneously well on the spelling test. The range of their scores is smaller than the range of scores of the entire second grade. Notice that the reliability coefficient for this group is lower (reliability = .27) than the coefficient for the entire grade.

This example illustrates one reason why it is incorrect to say "the test is reliable" or to say "the test is not reliable." As Gronlund and Linn (1990) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself. . . . Thus it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument." (p. 78, emphasis in original)

A test, in and of itself, cannot be reliable because reliability is not only a function of the items on a test, but also a function of who takes the test. As Rowley (1976) states, "It needs to be established that an instrument itself is neither reliable nor unreliable. . . . A single instrument can produce scores which are reliable and other scores which are unreliable" (p. 53).

Implications for Psychometrists, Therapists and Researchers

Education and psychology training programs teach students to read test manuals to examine reliability and validity coefficients. When evaluating norm statistics (such as reliability) reported in test manuals, psychometrists must be very careful that the intended examinee is similar to the normed group. As the example in Figure 1 illustrates, however, even when one cautiously selects a test that has yielded reliable scores for similar examinees in the past, it is incorrect to assume the test will yield reliable scores for all future uses of the test. A psychometrist is also cautioned to look at the homogeneity or heterogeneity of the norm group. For example, as the spelling test example has shown, if the norm group is incredibly heterogeneous compared to the group for whom the test is designed, one might expect that the reliabilities calculated for the intended groups will be lower than the ones reported in the manual. For this reason, and other reasons, it is important to calculate reliability *for every group to whom the test is given.*

Because reliability is a function of at least both the test and the test-takers, researchers as well as psychometrists should calculate reliability statistics on the scores of every group of persons that they measure. Researchers calculate reliability statistics on their own data for two reasons. The first reason to calculate reliability statistics on one's own data has been discussed previously: to cumulate evidence regarding the psychometric properties of measures. The second reason for calculating reliability statistics on one's own data is to determine the extent to which measurement error is limiting the effect sizes in the study of interest. Reinhardt (1991) cautioned researchers on this point, noting that "Prospectively, researchers must select measures that will allow detection of effects at the level desired; retrospectively, researchers must take reliability into account when interpreting findings" (p. 1).

An example using the effect size for a correlation coefficient explains this principle clearly. As Thompson (1991) explained, the correlation coefficient is the basis for all parametric statistics; "*all* classical analytic methods are correlational" (emphasis in original, p. 87). Therefore, the principle involving reliability coefficients and effect sizes has implications for *all* effect sizes in *all* common statistical procedures such as ANOVA, multiple regression, MANOVA, factor analysis, discriminant function analysis, and canonical correlation analysis. As Locke, Spirduso and Silverman (1987) noted, "the correlation between scores from two tests cannot exceed the square root of the product for reliability in each test" (p. 28). Written in equation form, the relationship between reliability and correlation looks like this:

$$r_{xy} < [(\textit{reliability of } X)(\textit{reliability of } Y)]^{.5}$$

This formula for the correlation between scores on X and scores on Y can be algebraically changed by squaring both sides of the equation to create an r^2 type of effect size. The new equation explains how reliability is related to effect sizes.

$$r^2_{xy} < [(\textit{reliability of } X)(\textit{reliability of } Y)]$$

The effect size can be no greater than the product of the reliability coefficients for the two measures that are being correlated.

An example illustrates how reliability influences effect size. One researcher was interested in the effects of self-esteem on achievement test scores. She used the Self-Esteem Scale of the Behavioral Assessment System for Children (BASC) (Reynolds & Kamphaus, 1992) to measure the self-esteem of third graders at a local elementary school. To measure achievement, she used the achievement scores from the standardized testing of the school district. She obtained two reliability coefficients, one for the self-esteem scores (r^{xx}= .60) and one for the achievement scores (r_{yy} = .90). Using the formula, $r^2_{xy} < [(\textit{reliability of } X)(\textit{reliability of } Y)]$, the researcher learned that the maximum effect size that she can obtain when she correlates self-esteem scores with achievement test scores is .54 (effect size < [(.6)(.9)]=.54). The researcher in this hypothetical example obtained an effect size of .52. Her uninformed colleague told her that the effect size was only "moderate." She replied to the colleague, "Moderate? How can you say that it is 'moderate' when the maximum effect size I could have found was .54? This is not a 'moderate' effect size. In the context of what could be (maximum =.54), .52 is a rather strong effect size." Thompson (1994) warned the would-be researcher to weigh the effects of reliability on effect sizes when planning and evaluating research:

> The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. For example, we may conduct studies that could not possibly yield noteworthy effect sizes given that score reliability inherently attenuates effect sizes. Or we may not accurately interpret the effect sizes in our studies if we do not consider the reliability of the scores we are actually analyzing. (p. 840)

Problems with True Score Model Estimates of Measurement Error

Now that the classical test theory reliability estimates have been defined and factors that influence them have been explained, it is important to illustrate some of the limitations of these estimates. One definition that lends itself particularly well to graphic illustration of these limitations is a definition offered by Crocker and Algina (1986). The reliability coefficient is "the proportion of observed score variance that may be attributed to variation in the examinees' true scores" (p. 116). An example has been drawn in Figure 3. The outer rectangle represents the observed score variance. The shaded area of the observed score variance is the true score variance (or 90% of the observed score variance). The unshaded portion of the observed score variance is the measurement error variance (10% of the observed score variance). Notice how the reliability coefficient (r_{xx} =.9) defines how much of the observed score variance is measurement error variance. Recall that in classical test theory, reliability is defined by the source of error being estimated (occasions, raters, forms or items).

Figure 3. Outer Rectangle = Observed Score Variance. Shaded Portion = True Score Variance. Unshaded Portion = Measurement Error Variance

From this diagram and the calculations that are used to estimate reliability in classical test theory, it becomes apparent that classical test theory only allows for the estimation of *one* type of error *at a time*--e.g., only inconsistencies across forms, but not across raters, items or occasions (Webb, Rowley, & Shavelson, 1988). It does not allow for estimations of *simultaneously*-occurring, *multiple* sources of measurement error. This major flaw in the true score model causes problems in the everyday use of reliability coefficients that have been derived using classical methods.

For example, a school psychologist is testing a boy to see if the boy will be given a mental retardation diagnosis. If the boy is given the diagnosis, then he will carry that diagnosis for many years to come. Therefore, it is very important that the tests that determine whether or not the boy receives a diagnosis yield scores that tend to have very high stability coefficients. That is to say, if the tests diagnose him as mentally retarded today, the tests should diagnose him as mentally retarded at many points in the future, because he will be carrying this label for many years. Would it not also be important to evaluate simultaneously whether or not the items on the tests enable the test administrator to generalize results on these tests to the greater domain of mental retardation (i.e., that the tests yield scores that tend to have high internal consistency coefficients)? Certainly, it would be important to ensure stability *and* internal consistency for tests with such far reaching implications. Classical test theory does not permit the researcher, psychometrist, or teacher to evaluate *simultaneously* the effects of both of these possible sources of error on examinees' observed scores. Therefore, in Figure 3 the portion of the diagram that represents error variance can represent only one source of error at a time and does not meet the needs of most researchers, teachers or psychologists (Webb, Rowley, & Shavelson, 1988).

There is a second problem with this limited model of measurement error (i.e., the true score model). The true score model does not account for error variance that may be caused by *interactions* between the different components of measurement error. Consider for example a testing scenario in which two judges assign ratings to candidates for entry into a graduate program based on 10 criteria. Table 10 outlines the 10 criteria and the ratings of the two judges on one of the applicants.

Table 10
Example of an Interaction Effect that is not Detected Using Classical Test Theory Reliability Estimates

Criteria	Judge # 1's Ratings	Judge # 2's Ratings	Total Score
Criterion 1	0	1	1
Criterion 2	0	1	1
Criterion 3	0	1	1
Criterion 4	0	1	1
Criterion 5	0	1	1
Criterion 6	1	0	1
Criterion 7	1	0	1
Criterion 8	1	0	1
Criterion 9	1	0	1
Criterion 10	1	0	1
	Total Rating = 5	Total Rating = 5	Total Score = 10

Notice that the judges both gave the candidate a total score of 5. According to classical test theory, the candidate's score is consistent across raters; thus, the inter-rater reliability coefficient seems to be high. Also notice that the candidate received a "1" on all of the criteria. Therefore, according to the true score model, the candidate's scores seem to be consistent across items and thus, internal consistency reliability is high. However, the true score model does not detect the obvious item-by-rater interaction effect. Such interactions can occur in common measurement situations, and can create sizeable and additional unique measurement error components.

Thus, the true score model has two serious shortcomings. The true score model is neither able to evaluate *simultaneously* two or more sources of measurement error, nor can it evaluate the *interaction effects* between different sources of measurement error. There is a method of deriving reliability estimates, however, that can detect more than one source of measurement error at a time as well as detect interaction effects between two sources of error. This method is known as generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability theory subsumes classical test theory as a special case (Thompson, 1991). A short description of generalizability theory follows. For more information, the interested reader is directed to Shavelson and Webb (1991). A note from Jaeger (1991), gives a flavor of the thoughts on generalizability theory in comparison to classical test theory: "Thousands of social science researchers will no longer be forced to rely on outmoded [classical theory] reliability estimation procedures when investigating the consistency of their measurements" (Jaeger, 1991, p. *x*).

Generalizability theory proceeds in two stages: first, a generalizability study (G-study) is performed, and then a decision study (D-study) is undertaken. In the G-study, the researcher estimates the amount of measurement error variance that is attributable to all sources of interest. Such sources of measurement error are the same ones used in classical test theory, such as inconsistencies in items, raters, forms or occasions. A G-study, however, measures these sources of measurement error, called facets, simultaneously. Therefore, it is possible to consider interaction effects between two sources of measurement error.

The term *facet*, which is used to describe each source of measurement error, can be illustrated using an analogy to the ANOVA procedure. Each facet can be conceptualized as a "way" or "main effect." Like "ways," facets have levels. The levels in each facet are called *conditions*. For example, a researcher wants to know how inconsistencies (sources of measurement error) in a test given on three occasions and with two forms may be affecting observed scores. Using ANOVA terminology, the researcher has a two-way design, the first way has three levels (the three different occasions on which the test was given), and the second way has two levels (the two different forms that were used). Using G-study terminology, the same example has a facet for occasions with three conditions and a facet for form which has two conditions.

Another G-study term is known as a *universe score*. A universe score is a hypothetical score (much like the true score in classical test theory). A universe score is an idealized score that tells a person's average score over all the possible items that could be asked about a particular subject, on all the possible occasions that the items could be administered, and across all possible forms. The variance that is attributed to the universe score is also referred to as the variance that is attributed to the *object of measurement*. (The object of measurement in education and psychology is usually a person or a "subject.") The object of measurement is also considered to be a *facet*, just as each of the sources of measurement error is considered to be a facet. Extending the ANOVA analogy, in a G-study there are two kinds of facets (ways). The first type of facet refers to a specific source of measurement error. The second type of facet refers to the object of measurement (Webb, Rowley & Shavelson, 1988).

A G-study actually uses ANOVA to partition the variance of the observed scores into the facet for the object of measurement and the facets for the different sources of measurement error. Said differently, using ANOVA terminology, the observed scores over many forms, occasions, raters, and items are the dependent variable scores. The independent variable scores, or the ways, are the object of measurement and the sources of measurement error. Thus, G-studies, like ANOVA, can yield information about interaction effects such as the item-by-rater interaction that was demonstrated in Table 10. Recall that this interaction was not detected by the classical test theory reliability estimates.

The second stage of analysis in generalizability theory is the decision study (D-study). The D-study uses information from the G-study to make decisions about the best measurement design. The D-study answers the question, "Which method of testing will minimize the amount of error variance in our study and will also be most efficient?" (Eason, 1991).

In summary, generalizability theory enables researchers, teachers and psychologists to do things that they would not be able to do using classical test theory methods. Generalizability theory investigates more than one source of measurement error at a time. Generalizability theory also allows for the investigation of interaction effects which are not detectable by classical test theory methods. Furthermore, generalizability theory reliability estimates yield information that helps teachers, researchers and psychologists to make the best decision about how to measure a content domain.

Conclusion

The present paper has demonstrated that several factors influence reliability coefficients as derived using classical test theory. While the qualities of a test do contribute to the magnitude of the reliabilities of the scores that the test yields, the qualities of that test certainly do not control whether or not all scores on the test can be called "reliable." Other factors including homogeneity of examinees, ability level of examinees vis-à-vis the test items, score variance, and test time-limits all have the potential to greatly influence reliability of scores. Therefore, it was first recommended that teachers, psychologists, and researchers calculate reliability estimates for each of their data sets, and not simply rely on reliability coefficients that are found in test manuals. Secondly, it was recommended that teachers, psychologists, and researchers refrain from saying and writing "the test is reliable," as such statements give the impression that reliability estimates do not change from one testing to the next. As Thompson has noted,

> This is not just an issue of sloppy speaking [or writing]--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice.
> (Thompson, 1992, p. 436)

Finally, some of the limitations of classical test theory reliability estimates were demonstrated, and a more advanced method of deriving reliability estimates using generalizability theory was described. Because these limitations can produce detrimental effects, it is recommended that psychologists, teachers, and researchers learn and use generalizability theory for deriving their reliability estimates.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J., (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of measurements.* New York: John Wiley.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Gronlund, N. E. (1976). *Measurement and evaluation in teaching* (3rd ed.). New York: Macmillan.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). *Applied statistics for the behavioral sciences.* Boston: Houghton Mifflin.

Jaeger, R. (1991). Foreword. In R. J. Shavelson & N. M. Webb, *Generalizability theory: A primer* (pp. *ix-x*). Newbury Park, CA: SAGE Publications.

Joint Committee on Standards for Educational Evaluation. (1994). The program evaluation standards: How to assess evaluations of educational programs. Thousand Oaks, CA: Sage.

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (1987). *Proposals that work: A guide for planning dissertations and grant proposals* (2nd ed.). Newbury Park, CA: Sage.

Reinhardt, B. M. (1991, January). *Factors affecting coefficient alpha: A mini Monte Carlo study.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 327 574)

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavioral Assessment System for Children.* Circle Pines, MN: American Guidance Service, Inc.

Rowley, G. L. (1976). The reliability of observational measures. *American Education Research Journal, 13,* 51-59.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5,* 417-426.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24,* 80-95.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434-438.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837-847.

Thompson, B., & Crowley, S. (1994, April). *When classical measurement theory is insufficient and generalizability theory is essential.* Paper presented at the annual meeting of the Western Psychological Association, Kailua-Kona, HA. (ERIC Document Reproduction Service No. ED 377 218)

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21,* 81-90.

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◇◇

Please enter a subscription to Research in the Schools for:

Name: ______________________________

Institution: ______________________________

Address: ______________________________

____

			COST
Individual Subscription ($25 per year)	Number of years	_____	_______
Institutional Subscription ($30 per year)	Number of years	_____	_______
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	_____	_______
TOTAL COST:			_______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
RESEARCH IN THE SCHOOLS
University of Alabama at Birmingham
School of Education, 233 Educ. Bldg.
901 13th Street, South
Birmingham, AL 35294-1250

Please note that a limited number of copies of Volume 1 are available and can be purchased for the same subscription prices noted above.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

* * *

MSERA Membership/Renewal Form

(Please print or type)

NAME: ____________________

TITLE: ____________________

INSTITUTION: ____________________

MAILING ADDRESS: ____________________
____
____

PHONE: ____________________ FAX ____________________

ELECTRONIC MAIL ADDRESS: ____________________

MSERA MEMBERSHIP: New___ Renewal ___
ARE YOU A MEMBER OF AERA? Yes___ No___
WOULD YOU LIKE INFORMATION ON AERA MEMBERSHIP? Yes___ No___

DUES:	Professional	$15.00	______
	Student	$10.00	______
VOLUNTARY TAX DEDUCTIBLE CONTRIBUTION TO MSER FOUNDATION			______
TOTAL			______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. Gerald Halpin
Auburn University
4036 Haley Center
Auburn, AL 36849

RESEARCH IN THE SCHOOLS
Mid-South Educational Research Association
and the University of Alabama at Birmingham
901 South 13th Street, Room 233
Birmingham, AL 35294-1250

BULK RATE
U.S. POSTAGE
PAID
PERMIT NO. 1256
BIRMINGHAM, AL

TM 0116

MSERA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and the University of Alabama at Birmingham.

Volume 4, Number 1 **Spring 1997**

Association Between Metals and Cognitive, Psychosocial, and Psychomotor Performance: A Review of the Literature 1
Richard F. Ittenbach, Collin Billingsley, Rebecca M. Spencer, John P. Juergens, Dennis A. Frate, and William H. Benson

Support for Prayer in the Schools Among Appointed and Elected Alabama Superintendents and School Board Members 9
Elizabeth M. Lacy and J. Jackson Barnette

Written Expression Reviewed 17
Jason C. Cole, Kathleen A. Haley, and Tracy A. Muenz

Assessing School Work Culture 35
William L. Johnson, Karolyn J. Snyder, Robert H. Anderson, and Annabel M. Johnson

Suspensions of Students With and Without Disabilities: A Comparative Study 45
Daniel Fasko, Deborah J. Grubb, and Jeanne S. Osborne

The Effects of Specific Interventions on Preservice Teachers' Scores on the National Teacher Exam 51
Hunter Downing, Sue Austin, Eileen Lacour, and Nancy Martin

Assistant Principals' Concerns About Their Roles in the Inclusion Process. 57
Louise L. MacKay and Patricia D. Burgess

Proper Use of the Two-Period Crossover Design When Practice Effects are Present 67
M. Suzanne Moody

James E. McLean and Alan S. Kaufman, Editors

469

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250**. **Please direct questions to jmclean@uab.edu**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1997 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 4, Number 1 **Spring 1997**

EDITORS
James E. McLean, *University of Alabama at Birmingham*
and Alan S. Kaufman, *Yale University School of Medicine*

PRODUCTION EDITOR
Margaret L. Rice, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Gypsy A. Abbott, *University of Alabama at Birmingham*
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Rhode Island*
Donald F. DeMoulin, *University of Memphis*
R. Tony Eichelberger, *University of Pittsburgh*
Daniel Fasko, Jr., *Morehead State University*
Ann T. Georgian, *Hattiesburg (Mississippi) High School*
Tracy Goodson-Espy, *University of North Alabama*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
JinGyu Kim, *National Board of Educational Evaluation (Korea)*
Jwa K. Kim, *Middle Tennessee State University*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Jerry G. Mathews, *Idaho State University*
Peter C. Melchers, *University of Cologne (Germany)*
Claire Meljac, *Unité de Psychopathologie de l'Adolescent (France)*
Soo-Back Moon, *Catholic University of Hyosung (Korea)*
Arnold J. Moore, *Mississippi State University*
Thomas D. Oakland, *University of Florida*
William Watson Purkey, *The University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Georgia Southern University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Valdosta State University*
Scott W. Snyder, *University of Alabama at Birmingham*
Bruce Thompson, *Texas A & M University*

GRADUATE STUDENT EDITORIAL BOARD
Margery E. Arnold, *Texas A & M University*
Vicki Benson, *The University of Alabama*
Alan Brue, *University of Florida*
Sue E. Castleberry, *Arkansas State University*
James Ernest, *University of Alabama at Birmingham*
Robin A. Groves, *Auburn University*
Harrison D. Kane, *University of Florida*
James C. Kaufman, *Yale University*
Sadegh Nashat, *Unité de Psychopathologie de l'Adolescent (France)*
Michael D. Scraper, *Kansas Newman College*
Sherry Vidal, *Texas A & M University*

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 1-8

Association Between Metals and Cognitive, Psychosocial, and Psychomotor Performance: A Review of the Literature

Richard F. Ittenbach and Collin Billingsley
The University of Mississippi

Rebecca M. Spencer
Vardaman Elementary School

John P. Juergens, Dennis A. Frate, and William H. Benson
The University of Mississippi

The detrimental effects of prolonged childhood exposure to toxic metals has received much attention in recent years. The purpose of the present paper is to review and synthesize findings pertaining to the association between metals (metallic compounds), particularly those involving lead, mercury, and cadmium, and the cognitive, psychosocial, and motoric functioning of school-age children and youth. Implications for research and primary, secondary, and tertiary interventions are discussed.

Knowledge of the detrimental effects of toxic metals began centuries ago. As early as 200 B.C., the philosopher Nicander found that exposure to high levels of lead could result in illness. Lead used as a sweetening agent and the storage of wine in leaden casks were two of the earliest culprits responsible for illness. Illnesses to persons of certain occupations have also been documented. For example, there are countless stories of the mirror makers of Venice and the hatters of London debilitated by long-term exposure to mercury. Many 18th century painters of Paris even fell victim to the lead-oxide in their paints (Fein, Schwartz, Jacobson, & Jacobson, 1983; Needleman, 1992; Weiss, 1983).

Lead poisoning of children was not formally documented nor investigated until the end of the nineteenth century. Studies by Turner (1897) and Gibson (1904), the earliest relevant studies known to date, found that a substantial number of children admitted to a hospital in Australia developed lead poisoning shortly after moving to a new house. A few decades later, Byers and Lord (1943) found that of 20 children previously treated for lead poisoning, 19 went on to develop academic or behavioral problems. Yet, it was not until Needleman et al.'s (1979) investigation into the effects of lead on the psychological and behavioral functioning of school children that researchers began to take seriously the role of metals as neurotoxins, potentially effecting the diminished academic abilities of children.

Throughout the past fifteen years, investigators have continued to reveal a relationship between metals and a threatened health status among children, adolescents, and adults. This relationship between science and practice is well illustrated by the recent history of lead in which maximum allowable exposure rates have fallen consistently from a high of 60 µg/dl (micrograms per deciliter of blood) in the 1960s to 10 µg/dl in 1991, with the added recommendation that all children under the age of two be tested for lead poisoning (Needleman, 1992). According to the National Health/Education Consortium, between 3 and 5 million (7%) children in the United States are exposed to lead concentrations high enough to be considered dangerous and potentially to cause neurocognitive

---

An earlier version of this paper was presented at the 1994 annual meeting of the Mid South Conference on Psychology in the Schools, Huntsville, Alabama. All correspondence regarding this manuscript should be addressed to Collin Billingsley, Ph.D., 5507 Waterpointe Cove, Tupelo, MS, 38801 (601-844-2544), or by E-mail: pymcb@sunset.backbone.olemiss.edu. A special note of thanks is extended to Jimmy Allgood, Julie Frate, and Jim O'Neal for their assistance on the broader research project. Richard F. Ittenbach, Ph.D., is an Associate Professor of Educational Psychology at The University of Mississippi. Collin Billingsley, Ph.D., is an Assistant Professor of Educational Psychology at Northeast Mississippi Community College. Rebecca M. Spencer, Ed.S., is the Principal at Vardaman Elementary School in Vardaman, MS. John P. Juergens, Ph.D., is a Research Associate Professor of Pharmacy Administration at The University of Mississippi. Dennis A. Frate, Ph.D., is a Research Professor of Pharmacy Administration and Anthropology at The University of Mississippi. William H. Benson, Ph.D., is a Professor of Pharmacology and Director of the Research Institute of Pharmaceutical Sciences at The University of Mississippi.

and neurobehavioral problems (Needleman, 1992). Preuss (1993) suggests that the rate of incidence may be as high as 10% to 50%.

According to Weiss (1983), adverse health effects should be pursued based on how people feel and operate and not simply based on indices of death or overt physical damage. While lead continues to be implicated as a contributor to the poor academic readiness and functioning of many children, the extent to which lead actually slows the acquisition of important academic skills remains uncertain. Less well known are the effects of other, and perhaps equally, potentially hazardous heavy metal elements such as mercury and cadmium.

Given the well-documented history of a) the relationship between environmental metal toxins and living organisms, b) the subtle and insidious means by which these neurotoxins may affect children's cognitive, psychosocial, and motor abilities, and c) the near void of literature review, a sample review of the respective literature seems long overdue. The purpose of the present paper is to review and synthesize findings in the literature pertaining to the association between metals, particularly lead, mercury, and cadmium, and the cognitive, psychosocial, and psychomotor functioning of school-age children and youth.

Criteria for Inclusion

Seven electronic data bases were accessed to identify relevant articles, monographs, and theses across the life, natural, and social science disciplines: ERIC, Dissertation Abstracts, Medline, PsycLit, Social Issues Resources Series, Social Sciences Index, and Toxline. Searches were conducted for periodicals up through mid 1995. In order to make inclusion for review, the articles must have met or been relative to each of the following four criteria: (a) human presence of organometallic elements (e.g., lead [Pb], mercury [Hg], aluminum [Al], cadmium [Cd]); (b) dependent measure(s) of cognition, academic achievement, psychosocial functioning, or psychomotor functioning; (c) children or youth 0 through 21 years of age; and (d) some indication of a toxicity.

Metals and Performance

Cognitive Performance

Few metals have commanded the attention of lead when it comes to body burden (hair-trace elements, dentine levels, and blood levels) and its effect on intellectual skills. For example, Ittenbach, Spencer, Juergens, Frate, and Benson (1995) found that 92% of the published studies involving children and metals as either independent or dependent variables involved lead. Results of the following published studies varied by age.

Needleman and his colleagues (1979) are generally cited as providing the seminal work in this area. These authors found, with children ages six to eight, inverse and statistically significant relationships between the amount of lead in a child's dentine and measures of general intelligence, verbal intelligence, and auditory processing ability. An 11-year follow-up study by Needleman, Schell, Bellinger, Leviton, and Allred in 1990 found that deficits in cognitive ability identified in the primary school years were still present in young adulthood. Specifically, those with high lead levels were found to have increased rates of school absenteeism, lower vocabulary and grammatical-reasoning scores, longer reaction times, and markedly higher odds of dropping out of high school.

Information obtained from infants using measures of pre- and post-natal blood lead have revealed an equally interesting pattern of results. It has been found that infants with relatively high levels of exposure to lead *in utero*(10 µg/dl <) had significantly lower indices of later mental development than infants with low levels of exposure. It has also been found that infants from lower socio-economic backgrounds evidenced greater deficits than infants from higher socio-economic backgrounds (Bellinger, Leviton, Waternaux, Needleman, & Rabinowitz, 1987, 1989). However, not everyone agrees with the direction of the evidence, particularly when it comes to school-age children. While the Cincinnati Lead Study team found lower measures of general and perceptual intelligence (approximately .5σ) for a sample of young urban children with high blood lead levels, differences disappeared when maternal intelligence and home environment were taken into account (Dietrich, Berger, Succop, Hammond, & Bornschein, 1993; Dietrich, Succop, Berger, & Keith, 1992).

Ernhart (1995) has proposed strong argument against the empirical literature's findings and sample reviews of this literature which suggests an association between low-level lead exposure and impaired cognitive-behavioral functioning. Her opposition is based upon proposed methodological and statistical flaws, such as a) asserting that the meta-analyses have demonstrated a spurious approach of selecting [nonexperimental] studies which utilized multiple measures of both risk factors and outcome variables; b) evading or not adjusting for various potential confounds (e.g., limited surrogates for nutrition, paternal intelligence, poverty, childhood illness, abuse and neglect, parental use of drugs); c) inconsistencies among intelligence scores and using measures other than full scale IQ scores across studies; d) errors in measurement were totally dismissed in most analyses of estimated lead effects; and e) effect modification (interaction effects between low birth weight and SES on the effects of lead level exposure, for example) is not replicated across studies. Ernhart (1985) also performed a much earlier meta-analysis of lead-exposure and its association with cognitive outcomes, reading tests, and behavior rating scales. In that reanalysis the author posited that the integrity of the lead-developmental outcome studies selected for meta-analysis and the approach of organizing and analyzing existing data were questionable.

Additionally, Pocock, Smith, and Baghurst (1994) performed a review of the epidemiological data in order to quantify the magnitude of the relationship between full scale IQ and body burden of lead in children five years of age or older. The authors concluded that low level lead exposure may cause slight cognitive impairment and, as reported by Ernhart, similar concerns were voiced regarding potential effects of extraneous variables and methodological limitations (observational research) upon both internal and external validity. However, Pocock et al.'s evidence revealed a slight, but potentially very important lead exposure effect associated with IQ deficit. Needleman (1993) responded to Scarr and Ernhart's (1993) study in opposition to negative effects by performing reanalyses of the published data and having independent reviewers perform reanalyses. He has refuted charges of scientific misconduct.

While there appear to be well documented discrepancies within the literature, the findings suggesting less conclusive or no evidence of association between low-level lead exposure and impaired cognitive functioning are of a methodological vein that appears to discount the respective practicality of effect size. That is, how strong in magnitude does the effect size need to be to be worthy of considerable practical importance; a question, in this case of human study, that cannot be easily answered by asserting an experimental level of methodological control and statistical probability alone. Significant statistical tests do not necessarily imply large effects, nor do they necessarily imply small effects. At least, sufficient evidence appears to exist to support the assumptions, upon which the clinical recommendations discussed in the conclusions of this literature review are based.

Laughlin's (1995) commentary on "a new" approach for the study of neurotoxicity compliments and reinforces Bellinger's (1995) comments, and it delivers refreshing promise for researching these important questions by utilizing the "experimental system." Bellinger reviews and suggests continued use of animal studies. He reviewed a primate study with monkeys in which a high level of experimental control was afforded and conclusions indicated chronically lead-exposed effects were associated with disruptions of certain nonplay behavior in monkeys at younger ages. Bellinger suggested that the discrepancies that exist among the animal data might also enhance our understanding of the mechanisms of effect modification upon human epidemiological studies.

Data obtained from the National Health and Nutrition Examination Survey and the 1980 census were analyzed by Sacks and Binder (1990) and produced an estimate of IQ points lost due to low-level lead exposure that exceeded an estimated 15.7 million raw score points. Schell's (1990) longitudinal follow-up study of the Needleman et al. (1979) study reported that exposure to lead was a significant factor in lowered reaction time, lowered IQ, and poor classroom behavior; three factors that were also found to be significant predictors of poor reading achievement. In light of these and other findings, Needleman (1992) has reported that children with unacceptably high levels of blood lead were seven times less likely to graduate from high-school and up to six times more likely to have a reading disability. Furthermore, Thatcher, Lester, McAlaster, Horst, and Ignasias (1983) found consistent and statistically significant relationships between hair lead and intelligence scores in children (ages 5 to 16) of average and gifted ability levels.

Within the context of special education specifically, lead has been implicated in a host of learning disabilities. Marlowe, Cossairt, Welch, and Errera (1984) found that when combined with aluminum or mercury, the negative effect of lead was increased beyond what it was believed to be capable of alone. A similar synergistic effect was found between lead and cadmium in which doses much lower than those that typically produce maladaptive symptoms may produce symptoms of mental retardation and borderline intelligence (Marlowe, Errera, & Jacobs, 1983). Additionally, Phil, Parkes, and Stevens (1979, cited in Marlowe et al., 1984) have linked lead and cadmium to learning disabilities and mental retardation. An interaction effect was also observed between lead, arsenic, and mercury in a sample of children with mental retardation living in rural Tennessee.

Psychosocial Behavior

Most of the psychosocial studies to date are based on anecdotal behavioral observations, formal ratings of behavioral observations, and correlations of blood, hair, and dentine levels with indices from various social-skills rating scales. Generalizations to more internal and less social areas of functioning are often made based on the magnitude of social deficit rather than a verified hypothesis about intra-personal phenomena. Bryce-Smith (1986) contends that the notion underlying Social Causation Theory, that neural networks responsible for social functioning are insulated from pathways pertaining to purely physiological functioning, is both fallible and misleading. Just as people vary in their response to alcohol and other voluntary intoxicants, they also vary in their response to involuntary toxicants such as heavy metals.

Marlowe and Errera (1982) found an association between the number of behavioral deficits (e.g., hyperactivity, distractibility, impulsiveness, short attention span, acting out, immaturity, disturbed peer relations, aggression, and mental disorders) and low levels of lead. In the Marlowe and Errera study, six of ten checklist items pertaining to disturbed peer relations correlated with lead level (cf. Albert et al., 1974). That is, children previously identified as having behavioral problems had significantly higher lead levels (11.51 ppm) than children serving as controls (6.52 ppm). The 2:1 ratio of blood-Pb in favor of children with behavioral problems is highly similar to the 2:1 ratio found earlier and independently by Hansen, Christensen, and Tarp (1980) using a sample of children (13.99 ppm) residing in a psychiatric hospital in Denmark and a matched sample of normal controls (6.90 ppm). The ratio is, however, moderately similar to results obtained by Kracke (1982) with children diagnosed as psychotic (9.95 ppm), neurotic (14.65 ppm), and normal (5.82 ppm). The 2:1 ratio also holds for hair-Pb as well, with emotionally disturbed children reporting 15.31 ppm as compared with 8.64 ppm for normal controls (Marlowe, Errera, Ballowe, & Jacobs, 1983).

Young African-American children with blood levels in excess of 15 μg/dl were reported by Sciarillo, Alexander, and Farrell (1992) to evidence more maladaptive behaviors than children with lower levels of Pb. Children falling in the high-exposed group demonstrated stronger Internalizing and Externalizing scores on the Child Behavior Checklist, had total behavior problem scores that were on average 5.1 points higher than children in the low-exposed group, and were 2.7 times more likely to have total battery scores in the clinical range (> 90th percentile). Similarly, Marlowe and his colleagues (e.g., Marlowe, Schneider, & Bliss, 1991; Marlowe, Stellern, Moon, & Errera, 1985) found relationships between increased lead and cadmium levels and indices of emotional disturbance. When compared with a control group, lead and cadmium levels were significantly higher in a sample of children with emotional disturbances and higher still in a sample of children with emotional disturbances who were also violence prone. Relationships between heavy metals and emotional disturbance were further complicated by decreased levels of phosphorous. Two studies cited by Marlowe et al. (1991) also support these findings. For example, Walsh (1983) observed that a sample of juvenile delinquents had significantly higher levels of lead and cadmium than nondelinquent youths. Schauss (1981) found significantly higher levels of organic metal compounds ($2\sigma <$) among a sample of violence prone youth than among a sample of non-violence prone controls. Studies with adults have yielded similar findings among violent offenders (e.g., Phil, Ervin, Pelletier, Diekel, & Strain, 1982).

The body burden of such heavy metals as lead, cadmium, and aluminum is commonly known, but the outcome of unusually high levels of trace minerals has not been as widely studied. For example while aluminum has no known dietary value to humans, it is the third most common element in the earth's crust and has many avenues of access into the human body. Cooke and Gould (1991) cite Howard (1984) and Moon and Marlowe's (1986) work attesting to the link between high aluminum levels and problem behaviors. In a study involving six toxic metals and fourteen trace minerals, Marlowe and Bliss (1993) found unusually high levels of lead, iron, aluminum, molybdenum, and vanadium to be significant contributors to parent and teacher estimates of maladaptive classroom behaviors of young children. The relationships between the aforementioned heavy metals, other trace minerals, and violence prone behaviors were further complicated by decreased levels of lithium (Marlowe et al., 1991). Only one investigative team, Harvey, Hamlin, Kumar, Morgan, and Spurgeon (1988), reported nonsignificant findings for the coefficients of association between blood-Pb and outcome variables.

Psychomotor Functioning

Weiss (1983) emphasized that all metals can be toxic, even ones that are essential in small amounts. While the toxic effects of methyl-mercury have been known for centuries, two more recent outbreaks caused by contaminated food have given researchers insight into the health hazards posed by varying levels of exposure. Prolonged low-level exposure can cause extensive damage to the cerebral cortex. Low-level exposure to methyl-mercury produced motor disturbances, general cognitive deficits, emotional disturbances and decreased alertness (Uphouse, 1981). Raloff (1991) has reported that even at very low doses, when encountered by developing fetuses, methylmercury can cause psychomotor retardation such as delays in speech or walking, severe brain damage, or other birth defects. Males with mild prenatal exposure seem to be more susceptible to damage than females (Clarkson, 1992).

Marlowe, Folio, Hall, and Errera (1982) studied the relationship between low-level lead absorption in samples of persons with mental retardation. They evaluated the relationships among nutrient minerals, heavy metals, and increased lead burdens and found that persons with mental retardation had higher concentrations of hair-trace elements of lead than persons without mental retardation and that similar differences were identified across six nutrient minerals. Over the past several years, Marlowe and his colleagues (e.g., 1986, 1992; Marlowe, Errera, & Jacobs, 1983; Marlowe, Stellern, Moon, & Errera, 1985; Moon, Marlowe, Stellern, & Errera, 1985) have investigated the effects of mercury, cadmium and aluminum on specific aspects of psychomotor skills (fine-motor and gross-motor functioning) and behavioral functioning. These correlational studies have shown that prolonged exposure to low levels of lead, cadmium, mercury, and aluminum are each likely to have deleterious effects on visual-motor performance such as, spatial analysis, object discrimination, motor speed, gross behavior activity levels and motor coordination tasks. In a more recent study, increased levels of aluminum were also found to be associated with decreased gross-motor skills such as running speed and agility and upper-limb coordination (Marlowe, 1992).

Needleman et al. (1990) investigated the long-term effects of exposure to low doses of lead on childhood psychomotor performance. Longitudinal data were collected from 1975 through 1978 and again in 1988. Their findings suggest that increased lead levels during childhood, especially 20 ppm or higher of hair trace, were significantly associated with poorer hand-eye coordination, slower reaction times, and slower finger tapping speeds. Needleman (1992) discussed findings of both short-term and long-term effects of lead exposure along these lines. His research concluded that children from first grade through fifth grade with elevated lead levels and who were evaluated again at age 18 experienced more hyperactive behaviors and poorer performance on fine motor tasks. Harvey et al. (1988) assessed 201 inner city children on motor tasks and various other cognitive and neuropsychological tasks. Their performance findings also revealed a significant association between blood lead level and tests requiring motor skills.

Each of the above assessments of the effects of methyl-mercury, lead, cadmium and aluminum on psychomotor development and functioning suggests that excessive exposure to these metals is strongly associated with deleterious effects on psychomotoric functioning. These studies utilized various modalities for assessing body burden which would suggest that regardless of the medium of body contraction, the metals and/or compounds typically cross the blood-brain barrier often at potentially harmful levels.

Conclusions and Implications

This review of how the consequences of excessive childhood exposure to metals affect cognitive, psychosocial, and psychomotor levels of functioning clearly identifies the need for efficient, accessible, and effective community metal screening programs and further scientific research. There is at least one major methodological limitation preventing researchers from investigating these variables experimentally, the high risk of brain insult at certain blood levels. The observational and correlational research is well merited, because it affords more advanced knowledge of the intrusive negative effects of metals on cognitive, psychosocial and psychomotor development and functioning. Even if the metals affect only minimal brain/behavior insult, it is worthy of clinical assessment and future study due to potential "logarithmic" effects on respective areas of functioning.

Although the effects of lead are moderately documented and widely known with clearly defined endpoints at high concentrations, far less is known about other metals and trace elements, particularly the effects of long-term chronic exposure on the cognitive, psychosocial, and psychomotoric functioning of infants and children. It is important to understand the physiological and pharmacological implications of exposure to these compounds as they continue to accumulate in the environment.

For mental health care providers (e.g., psychiatrists, counselors, psychologists, social workers, nurses, and marriage and family therapists), a referral for a medical screening may be in order, contingent upon impaired psychological areas and their levels of impairment. If these levels of impairment are clinically significant (based on interviews, observations, formal instruments) and metal exposure is identified in the patient's history, the mental health professional may wish to inquire with public health officials as to the rate and likelihood of metal contamination in other members of the community. A clinical implication for all mental health service providers, especially those providing services to children and adolescents with any type of developmental, behavioral, or eating disorder, is the need to use intake assessment questions that allow for determining possible experiences with or exposure to chemicals of any kind. Questions such as, "Has your child ever swallowed common household chemicals or ingested any other nonnutritive substances on a regular basis?" should be asked. One may also ask about prior dwellings and any known environmental risks present in the respective neighborhoods. While this information is not necessary to establish etiology, it is most necessary to assist with evaluations, referrals, and recommendations for further action. Where contamination is realized and blood screenings produce positive readings for toxicity, interventions should be adapted to the child, school and community with caution. Possible interventions are discussed in the following paragraphs.

Contingent upon the deficit areas identified, direct remediation for the child might include both special education services in general academic areas (e.g., reading comprehension, reading recognition, and math reasoning) and neuropsychological rehabilitation for neurocognitive, neuromotor, and neurobehavioral deficits (e.g., dysphasia, constructional dyspraxia, sensory-perceptual deficits, visual-motor deficits, and attentional processing deficits). If the child displays excessive disruptive type behaviors, then psychological interventions might include a highly structured behavior management program involving all significant others and possibly individual psychotherapy to train adaptive stressor-coping skills and enhance interpersonal functioning.

For teachers and school administrators overlooking the deleterious effects of toxic metals, a child's academic aptitude and achievement may suffer to the point of hindering cognitive, social, and behavioral development. Although children living in poverty are not the only ones affected by toxic metals, children who live in impoverished environments certainly overrepresent the number of children who are exposed to metals and thus experience academic difficulty. It has been found that connections between children in poverty and increased rates of exposure to toxic metals are almost inescapable (Lyngbye, Hansen, Trillingsgaard, Beese, & Grandjean, 1990). If exposure to these metals, even in small amounts, produces the kinds of damage found by investigators reviewed in this and other papers, the social and economical cost of remediation, special education services, discipline and other school programs may be immense. Nevertheless, alerting school personnel to the potential ramifications of such contamination is essential, and programs designed to identify and ameliorate such potentially impairing conditions should be developed collaboratively and immediately by community mental health professionals and school professionals. Responding to the needs of school children and adolescents who are at risk of excessive exposure or who have been exposed to metals in sufficient quantities and are consequentially at risk for significant cognitive, psychosocial, and psychomotoric impairment remains a challenging if not daunting task. As mentioned above, school professionals (administrators, teachers, counselors) and mental health professionals should form a multi-disciplinary approach in order to effectively and efficiently deliver both the remediational and rehabilitative approaches suggested in the previous paragraph.

The task of identifying excessive metallic exposure is made particularly difficult because the effects of the aforementioned metals may be subtle, non-specific, and without clear endpoints, and may even pass as a host of other nonspecific disorders. Thus developing community-wide programs for prevention of exposure is difficult. Simply educating parents and community professionals with the information regarding factors that are highly suspect and known to increase risk of exposure is the most indicated method of prevention. These types of educational campaigns should be spearheaded by someone or some group (task force or social agency) within individual communities in order to implement a sound methodological approach and to accomplish the task effectively.

Finally, further enhancement of biopsychosocial development and functioning should occur if related questions such as the following are to be answered: How do we identify, define, and refine the health and academic outcomes of exposure to these metals? When identified, how do we intervene in the child's environment such that prevention, amelioration and remediation can be both effective and long-lasting--not just for the child with the presenting problem, but for others in the school or community as well? Finally, are our current methods of classifying mortality and disability data sensitive enough to identify links between exposure and mortality? These are some of the many questions educational and psychological researchers and service providers must ask themselves as we struggle to better meet the psychological, social, and physical needs of children living in an "environmentally aware" and highly industrialized society.

References

Albert, R. E., Shore, R. E., Sayers, A. J., Strehlow, C., Kneip, J. J., Pasternack, B. S., Friedhoff, A. J., Covan, F., & Cominio, J. A. (1974). Follow-up of children over-exposed to lead. *Environmental Health Perspectives*, 7, 33-40.

Bellinger, D. (1995). Interpreting the literature on lead and child development: The neglected role of the "experimental system." *Neurotoxicology and Teratology, 17,* 201-212.

Bellinger, D., Leviton, A., Waternaux, C., Needleman, H., & Rabinowitz, M. (1987). Longitudinal analyses of prenatal and postnatal lead exposure and early cognitive development. *The New England Journal of Medicine, 316*(17), 1037-1043.

Bellinger, D., Leviton, A., Waternaux, C., Needleman, H., & Rabinowitz, M. (1989). Low level lead exposure, social class, and infant development. *Neurotoxicology and Teratology, 10,* 497-503.

Byers, R. K., & Lord, B. E. (1943). Late effects of lead poisoning on mental development. *American Journal of Disabilities in Children, 66,* 471-483.

Bryce-Smith, D. (1986). Environmental chemical influences on behavior, personality, and mentation. *International Journal of Biosocial Research, 8*(2), 115-150.

Clarkson, T. W. (1992). Mercury: Major issues in environmental health. *Environmental Health Perspectives, 100,* 31-38.

Cooke K., & Gould, M. H. (1991). The health effects of aluminum--A review. *Journal of The Royal Society of Health, Vol.# III,* 163-168.

Dietrich, K. N., Berger, O. G., Succop, P. A., Hammond, P. B., & Bornschein, R. L. (1993). The developmental consequences of low to moderate prenatal and postnatal lead exposure: Intellectual attainment in the Cincinnati lead study cohort following school entry. *Neurotoxicology and Teratology, 15,* 37-44.

Dietrich, K. N., Succop, P. A., Berger, O. G., & Keith, R. W. (1992). Lead exposure and the central auditory processing abilities and cognitive development of urban children: The Cincinnati lead study cohort at age 5 years. *Neurotoxicology and Teratology, 14,* 51-56.

Ernhart, C. B. (1985). Subclinical lead level and developmental deficits: Re-analyses of data. *Journal of Learning Disabilities, 18,* 475-479.

Ernhart, C. B. (1995). Inconsistencies in the lead-effects literature exist and cannot be explained by "effect modification." *Neurotoxicology and Teratology, 17,* 277-233.

Fein, G. G., Schwartz, P. M., Jacobson, S. W., & Jacobson, J. L. (1983). Environmental toxins and behavioral development. *American Psychologist, 39,* 1188-1197.

Gibson, J. L. (1904). A plea for painted railings and painted walls of rooms as the source of lead poisoning among Queensland children. *Australia Medical Gazette, 23,* 149-153.

Hansen, J. C., Christensen, L. B., & Tarp, U. (1980). Hair lead concentrations in children with minimal cerebral dysfunction. *Danish Medical Bulletin, 27,* 259-263.

Harvey, P. G., Hamlin, M. W., Kumar, R., Morgan, G., & Spurgeon, A. (1988). Relationships between blood lead, behavior, psychometric and neuropsychological test performance in children. *Journal of Developmental Psychology, 6,* 145-156.

Howard, J. M. (1984). Clinical import of small increases in serum aluminum. *Clinical Chemistry, 30,* 1722-1723.

Ittenbach, R. F., Spencer, R. M., Juergens, J. P., Frate, D. A., & Benson, W. H. (1995). Metals and school learning: A review of investigative techniques. *Perceptual and Motor Skills, 81,* 1079-1090.

Kracke, K. R. (1982). Biochemical basis for behavioral disorders in children. *Journal of Orthomolecular Psychiatry, 11,* 289-293.

Laughlin, N. K. (1995). A new approach for the study of the neurotoxicity of lead. *Neurotoxicology and Teratology, 17,* 235-236.

Lyngbye, T., Hansen, O. N., Trillingsgaard, A., Beese, I., & Grandjean, P. (1990). Learning disabilities in children: Significance of low-level lead exposure and confounding factors. *Acta Paediatrica Scandanavia, 79,* 352-360.

Marlowe, M. (1986). *Exposure to metal pollutants and behavioral disorders in children: A review of the evidence.* Paper presented at the Annual Convention of the Council for Exceptional Children, New Orleans, LA. (ERIC Document Reproduction Service No. ED 268 742)

Marlowe, M. (1992). Low level aluminum exposure and childhood motor performance. *Journal of Orthomolecular Medicine, 7,* 147-152.

Marlowe, M., & Bliss, L. B. (1993). Hair element concentrations and young children's classroom and home behavior. *Journal of Orthomolecular Medicine, 8*(2), 79-88.

Marlowe, M., Cossairt, A., Welch, K., & Errera, J. (1984). Hair mineral content as a predictor of learning disabilities, *Journal of Learning Disabilities, 17*(7), 418-421.

Marlowe, M., & Errera, J. (1982). Low lead levels and behavior problems in children. *Behavior Disorders, 7,* 163-172.

Marlowe, M., Errera, J., Ballowe, T., & Jacobs, J. (1983). Low metal levels in emotionally disturbed children. *Journal of Abnormal Psychology, 93*, 386-390.

Marlowe, M., Errera, J., & Jacobs, J. (1983). Increased lead and cadmium burdens among mentally retarded children and children with borderline intelligence. *American Journal of Mental Deficiency, 87*(5), 477-483.

Marlowe, M., Folio, R., Hall, D., & Errera, J. (1982). Increased lead burdens and trace-mineral status in mentally retarded children. *The Journal of Special Education, 16*(1), 87-99.

Marlowe, M., Schneider, H. G., & Bliss, L. B. (1991). Hair mineral analysis in emotionally disturbed and violence prone children. *International Journal of Biosocial Medical Research, 13*(2), 169-179.

Marlowe, M., Stellern, J., Moon, C., & Errera, J. (1985). Main and interaction effects of metal pollutants on visual motor performance, *Archives of Environmental Health, 40*, 221-225.

Moon, C., & Marlowe, M. (1986). Hair-aluminum concentrations and children's classroom behavior. *Biological Trace Element Research, 11*, 5-12.

Moon, C. Marlowe, M., Stellern, J., & Errera, J. (1985). Main and interaction effects of metal toxins on cognitive functioning. *Journal of Learning Disabilities, 18*, 217-221.

National Health/Education Consortium. (1990). *Healthy brain development: Precursor to learning.* Washington, DC: National Commission To Prevent Infant Mortality. (ERIC Document Reproduction Service No. 329 345)

Needleman, H. L. (1992). *The poisoning of America's children: Lead exposure, children's brains, and the ability to learn.* Washington, DC: National Commission to Prevent Infant Mortality, National Health/Education Consortium. (ERIC Document Reproduction Service No. ED 354 071)

Needleman, H. L. (1993). "On being a whistleblower: The Needleman case:" Reply to Ernhart, Scarr, and Geneson. *Ethics and Behavior, 3*, 95-101.

Needleman, H. L., Gunnoe, C., Leviton, A., Reed, R., Peresie, H., Maher, C., & Barrett, P. (1979). Deficits in psychologic and classroom performance of children with elevated dentine levels. *New England Journal of Medicine, 300*(13), 689-695.

Needleman, H. L., Schell, A., Bellinger, D., Leviton, A., & Allred, E. N. (1990). The long-term effects of exposure to low doses of lead in childhood: An 11-year follow-up report. *New England Journal of Medicine, 322*(2), 83-88.

Phil, R. O., Ervin, R., Pelletier, G., Diekel, W., & Strain, W. (1982). Hair element content of violent criminals. *Canadian Journal of Psychiatry, 27*, 533-534.

Phil, R. O., Parkes, M., & Stevens, R. (1979). Nonspecific interventions with learning disabled individuals. In R. Knights and D. Bakker (Eds.), *Rehabilitation, treatment, and management of learning disorders.* Baltimore, MD: University Park Press.

Preuss, H. G. (1993). A review of persistent, low-grade lead challenge: Neurological and cardiovascular consequences. *Journal of the American College of Nutrition, 12*(3), 246-254.

Pocock, S. J., Smith, M., & Baghurst, P. (1994). Environmental lead and children's intelligence: A systematic review of the epidemiological evidence. *British Medical Journal, 309*, 1189-1197.

Raloff, J. (1991). Mercurial risks from acid's reign. *Science News, 139*, 152-156.

Sacks, J. J., & Binder, S. (1990). Points of potential IQ lost from lead. *Journal of the American Medical Association, 264*, 2212.

Scarr, S., & Ernhart, C. B. (1993). Of whistleblowers, investigators, and judges. *Ethics and Behavior, 3*, 199-206.

Schauss, A. G. (1981). Comparative hair mineral analysis results of 21 elements in a random selected behaviorally 'normal' 19-59 year old population and violent adult criminal offenders. *International Journal of Biosocial Research, 1*, 21-41.

Schell, A. (1990). *Low level lead exposure during childhood and reading achievement in childhood and young adulthood.* Unpublished doctoral dissertation, Boston University.

Sciarillo, W. G., Alexander, G., & Farrell, K. (1992). Lead exposure and child behavior. *American Journal of Public Health, 82*(10), 1356-1360.

Thatcher, R. W., Lester, M. L., McAlaster, R., Horst, R., & Ignasias, S. W. (1983). Intelligence and lead toxins in rural children. *Journal of Learning Disabilities, 16*(6), 355-359.

Turner, A. J. (1897). Lead poisoning among Queensland children. *Australia Medical Gazette, 16*, 475-479.

Uphouse, L. L. (1981). *Environmental effects on health with special emphasis on neurotoxicology.* Washington, DC: The Research Forum on Children and Youth. (ERIC Document Reproduction Service No. ED 214 638)

Walsh, W. (1983). Cited in Raloff, R. Locks (1983). A key to violence. *Science News, 20*, 122-124.

Weiss, B. (1983). Behavioral toxicology and environmental health science. *American Psychologist, 39*, 1175-1187.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 9-16

Support for Prayer in the Schools Among Appointed and Elected Alabama Superintendents and School Board Members

Elizabeth M. Lacy
Decatur City Schools, Alabama

J. Jackson Barnette
Independent Consultant

Seven school settings where prayer could be present were identified: beginning of the school day in classroom, beginning of school day over public address system, moment of silence, sporting events, baccalaureate service, graduation ceremony, and prayer clubs. Respondents were asked to indicate support for these as student-initiated and school-sponsored. Support among school decision-makers for school prayer was high. Support was higher for student-initiated prayer activities as compared with school-sponsored prayer, particularly for superintendents. Relative to school-sponsored prayer, board members were more supportive than superintendents. Relative to student-initiated prayer and school-sponsored prayer, elected decision-makers were more supportive than appointed decision-makers. Prayer decisions are influenced highly and about equally from Supreme Court decisions, laws, and community sentiment.

The available literature relating to the issue of public school prayer is plentiful and diverse. The majority of the early settlers of North America were motivated by a need for religious freedom rather than the religious tolerance to which most were subjected in their mother countries (Cubberley, 1962; Kelley, 1984). The first North American public schools were started in Massachusetts in 1647 for the purpose of teaching Bible reading (Thomas & Anderson, 1982). In a sustained reaction to the position of Rome which discouraged Bible reading by the laity, Protestant leaders in the colonies required Bible reading in the newly created schools (Pfeffer, 1975).

The Puritan influence throughout New England contributed greatly to the public educational system which was to follow and was generally adopted by other colonies (Cubberley, 1962). The somber Calvinistic influence produced a parochial school system in which children were surrounded daily with religious oversight. Indeed, one of the primary duties of schoolmasters or teachers was to "catechize their scholars in the principles of the Christian religion," and it was "a chief part of the schoolmaster's religious care to commend his scholars and his labors amongst them unto God by prayer morning and evening" (Cubberley, 1962, p. 41).

A system of free, tax supported schools existed by the 1850s in the northern United States; after the War Between the States, the southern states followed. Initial resistance to school taxation, usually in the form of property taxes, was successfully countered by reformers using the argument that the owners were considered "trustees" of the community's wealth (Nasaw, 1979). Compulsory attendance laws were passed and most public schools continue to require attendance until age 16 (Thomas & Anderson, 1982).

While all children were required to attend public school, they remained strongly Protestant in orientation. The curriculum which existed under the church school structure continued including reading, writing, arithmetic, and religion; additionally, many of the clergy remained as teachers (Pfeffer, 1975).

The Establishment and Free Exercise Clauses

Two of the primary clauses in the First Amendment are (a) Congress shall not establish a religion (Establishment Clause) and (b) Congress shall not forbid citizens to exercise freedom of choice in religion and worship (Free Exercise Clause). Many states refused to sign the Constitution without the promise that amendments would be added which would guarantee religious freedom and individual rights (Brown, 1983). The Fourteenth Amendment made these prohibitions applicable to the states in 1868 (O'Reilly & Green, 1992).

In deciding court cases relating to church and state matters in this century, the United States Supreme Court has relied heavily on the writings of James Madison and

---

Elizabeth Lacy is Principal of Cedar Ridge Middle School, Decatur, AL, and J. Jackson Barnette is an independent consultant in Tuscaloosa, AL. Please direct all correspondence to Jack Barnette at 2428 Brandon Parkway, Tuscaloosa, AL 35406, (205-349-4489). Email: jbarnett@dbtech.net.

Thomas Jefferson (Alley, 1988). Jefferson defined the First Amendment in 1802 as "building a wall of separation between church and state" (Alley, 1988, p. 3) and this concept was supported by Madison (Starr, 1985).

The Establishment Clause and the Free Exercise Clause may overlap in some cases; however, they prohibit two different kinds of governmental intrusion upon religious freedom (Douglas, 1966; Lowry, 1963). The Free Exercise Clause assures every person in the United States the right to worship in the manner he chooses or not to worship (Douglas, 1966). Additionally, the state may not do anything perceived as restraint, constraint, coercion, or compulsion as they relate to any man's attempt to practice or not to practice religion (LaNoue, 1967). Religious exercises in public schools may have inherent in them the element of compulsion for the child who wishes not to conform (Douglas, 1966).

Court Challenges

Despite restrictions which have been placed on public schools relating to the First Amendment, several court cases have arisen concerning prayer and Bible reading in public school classrooms. *Engle v. Vitale* (1962) and *Abington Township, Pennsylvania v. Schempp* (1963) were such cases (cited in Schamel & Mueller, 1989). In New York, the Regents prayer was prohibited, and in Pennsylvania, the practice of required group prayer was forbidden. Establishment Clause jurisprudence regarding school prayer since these Supreme Court decisions has been primarily based on the *Lemon v. Kurtzman* (1971) case which has evolved into a three-pronged test (cited in Starr, 1985). The first prong requires the state to have a secular purpose; the second prong requires that the primary effect not advance or inhibit religion; and the third prong says the state's actions must not foster excessive entanglement with religion (Barber, 1992).

Over 20 years after the original court cases involving public school prayer were decided, a conservative group called the Moral Majority advanced the argument for sanctioning religion in schools to counter the secular humanist, or enemy (Brown, 1983). Additionally, the 19th Annual Gallup Poll of the Public's Attitude Toward the Public Schools indicated 68% of the respondents favored an amendment to the United States Constitution that would allow public school prayer (Gallup & Clark, 1987). In a move toward accommodation, the Supreme Court in *Widner v. Vincent* (1981) decided that student prayer clubs should have the right to meet using public university facilities (cited in Sendor, 1983); additionally, this right was extended to high school students in *Bender v. Williamsport Area School District* (1986).

While school-sponsored prayer in public school classrooms was clearly forbidden by the Supreme Court decision *Wallace v. Jaffree* (1985), prayer during public school graduation was first addressed by the Supreme Court in *Lee v. Weisman* (1992) (cited in Rankin & Strope, 1994; Zirkel, 1994). Citing the Establishment Clause, the justices ruled that prayers delivered by clergy at public school graduation ceremonies were forbidden (Rankin & Strope, 1994). Questions about student-initiated prayer at graduation arose immediately, creating confusion for superintendents and other school officials (Rankin & Strope, 1994). Superintendents who represented school districts located in highly religious communities felt they would cause more problems changing the graduation format than following the Supreme Court ruling; indeed, some superintendents continued having student-selected ministers as guest speakers (Rankin & Strope, 1994).

To further complicate the issue of prayer at public school graduation, the Fifth Circuit United States Court of Appeals in *Jones v. Clear Creek Independent School District* (1991) stated these prayers were permissible based on four criteria: (1) the graduating seniors must choose the invocation and benediction, (2) the invocation and benediction must be delivered by a student volunteer, (3) if a prayer is given, it must be nonsectarian, and (4) the student volunteer must not proselytize (cited in Vacca & Hudgins, 1994). In allowing this decision to stand in June 1993, the Supreme Court permitted a defensible position for legal graduation prayers (Vacca & Hudgins, 1994). A clear distinction remains, however, between the constitutionality of student-initiated prayer and school-sponsored prayer based on the *Lee* and *Jones* decisions (Horner & Barlow, 1994).

United States Department of Education Principles

In response to the discussion of public school prayer and in an effort to communicate with all public school superintendents and school officials, U.S. Secretary of Education Richard W. Riley sent a letter and a list of principles to superintendents stating the obligations required by the First Amendment. Within the list of principles, the following tenants were presented:

1. Students may pray in public school in a nondisruptive manner, but this right does not include the right to have a captive audience listen.
2. School officials may not establish religious baccalaureate services or mandate prayer at graduation ceremonies.

3. Schools must be neutral toward religion; however, they may play an active role in teaching civic values and the moral code that provide community cohesiveness (U.S. Department of Education, 1995).

Clearly, public school decision-makers face difficult decisions and pressures from a wide variety of stakeholders in the school prayer controversy. This research was designed to assess support for student-initiated and school-sponsored prayer and to determine if support is related to position held (superintendent or school board member), method of selection for the position (appointed or elected), or both. School prayer is an emotional, highly political issue. Dealing with such an issue effectively requires an understanding of the attitudes of the various educational decision-makers involved. This study adds to the body of knowledge related to this issue and hopefully will permit various groups to have greater understanding of the attitudes of other decision-makers.

Purpose

The purpose of this research was to answer the following questions:

1. To what extent do superintendents and school board members support student-initiated and school-sponsored prayer?
2. Is support different between student-initiated and school-sponsored prayer and, if there is a difference, is that difference related to interaction or main effects of position and method of selection?
3. Is there an interaction of position (superintendent or school board member) and method of selection (appointed or elected) on support for student-initiated prayer, and, if there is no interaction, are there differences between the positions and are there differences between the methods of selection?
4. Is there an interaction of position (superintendent or school board member) and method of selection (appointed or elected) on support for school-sponsored prayer and, if there is no interaction, are there differences between the positions and the methods of selection?
5. How do respondents rate the influence of U.S. Supreme Court decisions, laws passed by the U.S. Congress and the Alabama Legislature, and community sentiment on their decision-making in the area of school prayer, and are these different between the positions and between methods of selection?

Methods

Sample

Respondents were solicited from two populations: Alabama public school superintendents ($N = 134$) and school board members ($N = 740$). The entire population of superintendents and a 20% random sample of school board members ($n = 148$) were solicited. Surveys were mailed to the 134 superintendents and 148 school board members, a total of 282 potential respondents. Useable surveys were returned by 104 superintendents, a return rate of 77.6% for superintendents, and by 106 board members, a return rate of 71.6% for board members. The overall return rate for all potential respondents was 74.5%.

Instrumentation

Seven settings where prayer could be present were identified: beginning of the school day in classroom, beginning of school day over public address system, moment of silence, at sporting events, at baccalaureate service, at graduation ceremony, and in prayer clubs. A survey was developed which asked respondents to indicate level of support for each of these if they were student-initiated or school-sponsored. In addition, respondents were asked to rate influence of Supreme Court decisions, laws passed by Congress and the Alabama Legislature, and community sentiment on their decision-making. The instrument was field tested with 12 former superintendents and 12 former school board members. Cronbach's alpha for the 17 returned instruments was .92. The field test resulted in some minor modifications in the wording of a few items and the response categories. On the final draft of the instrument, support was indicated using a five-point Likert scale of 1= strongly opposed to 3= neutral to 5= strongly support for the student-initiated and school-sponsored prayer items and a five-point Likert scale of 1= not important to 3= neutral to 5= very important for the four influence items. Cronbach's alpha for the final instrument for the final analysis sample was .91.

Analysis of Data

Analysis included both descriptive and inferential statistics. Descriptive statistics in the form of percentage of response were used to answer question 1, related to overall support for the prayer issues. Scores were determined for each subscale (student-initiated and school-sponsored) by averaging the seven item responses. Comparing the student-initiated level of support and school-sponsored means, and any possible relationship of the position and/or selection variables (question 2) was conducted using multivariate analysis of variance. Questions 3 and 4 were analyzed using univariate 2-way analysis of variance. When differences were found between the groups on the subscale means, follow-up was conducted using chi-square tests on an item basis. For question 5, overall percentage responses were presented for the four items and each of the four item frequency distributions were compared using chi-square tests between the two positions and two methods of selection. All significance tests used .05 for alpha. For the by-item comparisons within each question-based comparison, the level of significance was adjusted using a Bonferroni correction; alpha for significance was based on .05 divided equally across the number of items.

Results

Question 1 was: To what extent do superintendents and school board members support student-initiated and school-sponsored prayer? Two tables are used to present the results related to this question. Table 1 presents the results for the student-initiated aspects of school prayer. In general, the respondent group was supportive on all seven aspects. Percent support ranged from 58.4% to 86.1% for the seven items. Of these, the level of support was highest for baccalaureate service (86.1%), followed by: graduation ceremony (80.2%), prayer clubs (79.3%), moment of silence (77.8%), and sporting events (77.3%). Table 2 presents the results for the school-sponsored aspects. In general, there was support for all of these except for one. The highest levels of support were for: baccalaureate service (76.9%), graduation ceremony (73.0%), sporting events (68.3%), moment of silence (66.8%), and prayer clubs (66.7%). Of these items, one clearly had lower support; the setting of beginning of school day on the public address system had 25.3% strongly opposing this practice.

Table 3 presents the summary statistics for the two subscales and the difference between them (student-initiated minus school-sponsored) for all of the respondent group crossed categories. Question 2 was: Is support different between student-initiated and school-sponsored prayer and, if there is a difference, is that difference related to interaction or main effects of position and method of selection? The basis for answering this question was multivariate analysis of variance. As indicated in Table 4, there was a significant difference in the subscale measures, $F(1, 206) = 31.560$, $p < .05$. The mean for the student-initiated subscale ($M = 4.206$) was significantly higher than the mean for school-sponsored ($M = 3.853$). This difference in the subscale means was not related to the interaction of position and selection, $F(1, 206) = 0.053$, $p > .05$, nor by selection, $F(1, 206) = 1.788$, $p > .05$. However, the difference was related to position, $F(1, 206) = 6.450$, $p < .05$. There was a higher difference for the superintendent group ($M = 0.534$) as compared with the school board group ($M = 0.174$).

Table 1
Percent Response on Student-Initiated Prayer Items, Total Group

		Strongly Oppose		Neutral		Strongly Support
Item	*n*	1	2	3	4	5
Beginning the school day (classroom prayer)	205	12.7	5.4	15.1	10.2	56.6
Beginning the school day (prayer on public address)	202	16.8	8.4	16.3	12.9	45.5
Moment of silence (school-wide or by class)	203	4.9	3.4	13.8	18.7	59.1
Sporting events	207	7.7	2.9	12.1	12.6	64.7
Baccalaureate service	208	3.8	1.0	9.1	10.6	75.5
Graduation ceremony	207	5.3	2.4	12.1	8.7	71.5
Prayer clubs	208	4.8	1.9	13.9	7.7	71.6

Table 2
Percent Response on School-Sponsored Prayer Items, Total Group

		Strongly Oppose		Neutral		Strongly Support
Item	*n*	1	2	3	4	5
Beginning the school day (classroom prayer)	205	18.5	7.3	19.5	9.8	44.9
Beginning the school day (prayer on public address)	198	25.3	10.1	21.2	10.6	32.8
Moment of silence (school-wide or by class)	202	9.9	3.0	20.3	19.3	47.5
Sporting events	208	14.9	4.3	12.5	13.0	55.3
Baccalaureate service	208	10.1	2.9	10.1	9.6	67.3
Graduation ceremony	208	11.5	5.3	10.1	9.1	63.9
Prayer clubs	207	12.6	3.9	16.9	9.7	57.0

Question 3 was: Is there an interaction of position (superintendent or school board member) and method of selection (appointed or elected) on support for student-initiated prayer, and if there is no interaction, are there differences between the positions and between methods of selection? The means for the 2-by-2 structure are found in Table 3. These means were compared using analysis of variance, results of which are presented in Table 5. There was no significant interaction, $F(1,206) = 0.18$, $p > .05$. There was, however, a significant main effect for the method of selection variable, $F(1,206) = 14.24$, $p < .05$. The mean for the elected group ($M = 4.493$) was significantly higher than the mean of the appointed group ($M = 3.959$). As a follow-up of this difference, the item distributions of these two groups were compared using chi-square homogeneity of proportions tests, with alpha adjusted for the set of seven items. Results are presented in Table 6. There was higher support on every item for the elected group. Of these, three were statistically significant: beginning of school day in the classroom, beginning of school day over public address, and sporting events.

Table 3
Means and Standard Deviations by Position and Selection for Student-Initiated Prayer, School-Sponsored Prayer, and Difference

Position	Selection	*n*	Student-Initiated		School-Sponsored		Difference	
			M	*SD*	*M*	*SD*	*M*	*SD*
Superintendents	Appointed	70	3.864	1.024	3.285	1.257	0.580	1.058
	Elected	34	4.403	0.830	3.962	1.256	0.441	0.984
Board Members	Appointed	43	4.113	1.078	3.822	1.306	0.291	0.949
	Elected	63	4.542	0.563	4.447	0.606	0.094	0.375
Superintendents	Total	104	4.041	0.988	3.506	1.291	0.534	1.032
Board Members		106	4.368	0.835	4.194	0.997	0.174	0.673
Total	Appointed	113	3.959	1.047	3.489	1.297	0.470	1.023
	Elected	97	4.493	0.668	4.277	0.913	0.216	0.672
Total	Total	210	4.206	0.930	3.853	1.200	0.352	0.886

Table 4
Multivariate F Tests Comparing Student-Initiated and School-Sponsored by Position and Selection

Source	*df*	*F*	*p*
Measures	1, 206	31.560	<.001
Measures by Position	1, 206	6.450	.012
Measures by Selection	1, 206	1.788	.183
Measures by Position and Selection	1, 206	0.053	.818

Table 5
Analysis of Variance for Student-Initiated Prayer Subscale

Source	*df*	*MS*	*F*	*p*
Position	1	1.809	2.28	.133
Selection	1	11.310	14.24	<.001
Interaction	1	0.146	0.18	.669
Error	206	0.794		
Total	209			

Table 6
Comparison by Selection on Student-Initiated Prayer Items

Item	*n*	Appointed % support	Elected % support	χ^2	*p*
Beginning the school day (classroom prayer)	205	58.6	76.6	21.752	<.001*
Beginning the school day (prayer on public address)	202	50.0	68.1	15.744	.003*
Moment of silence (school-wide or by class)	203	73.6	82.8	6.356	.174
Sporting events	207	67.9	88.4	18.536	.001*
Baccalaureate service	208	81.3	91.7	11.511	.021
Graduation ceremony	207	73.2	88.4	13.189	.010
Prayer clubs	208	77.7	81.3	4.599	.331

*$p < .05$, adjusted alpha= .0071

BEST COPY AVAILABLE 484

Question 4 was: Is there an interaction of position (superintendent or school board member) and method of selection (appointed or elected) on support for school-sponsored prayer, and if there is no interaction, are there differences between positions and between methods of selection? The 2-by-2 cell means are found in Table 3. As indicated in Table 7, there was no significant interaction, $F(1, 206) = 0.03$, $p > .05$. There were differences for both main effects.

Table 7
Analysis of Variance on School-Sponsored Subscale

Source	*df*	*MS*	*F*	*p*
Position	1	12.626	10.18	.002
Selection	1	20.483	16.51	<.001
Interaction	1	0.033	0.03	.871
Error	206	1.241		
Total	209			

For the variable of position, $F(1, 206) = 10.18$, $p < .05$, the mean for the school board members ($M = 4.194$) was significantly higher than the mean for the superintendents ($M = 3.506$). Comparisons by item are found in Table 8. Board members had higher levels of support on all seven items. Four of the items had significant differences: beginning of school day in the classroom, beginning of the school day on the public address system, baccalaureate service, and graduation ceremony.

Table 8
Comparison by Position on School-Sponsored Prayer Items

Item	*n*	Supt. % support	Bd. Member % support	χ^2	*p*
Beginning the school day (classroom prayer)	205	37.0	71.4	27.315	<.001*
Beginning the school day (prayer on public address)	198	32.7	54.6	16.638	.002*
Moment of silence (school-wide or by class)	202	61.0	72.5	5.992	.200
Sporting events	208	59.6	76.9	10.699	.030
Baccalaureate service	208	66.0	87.6	16.916	.002*
Graduation ceremony	208	62.1	83.8	18.146	.001*
Prayer clubs	207	62.1	71.2	3.329	.504*

$p < .05$, adjusted alpha = .0071

Elected superintendents and board members were more supportive ($M = 4.277$) than appointed superintendents and board members ($M = 3.489$), $F(1,206) = 16.51$, $p < .05$. Comparison of school-sponsored items between methods of selection are found in Table 9. On all items, elected were more supportive than appointed. Five of these were significant: beginning of school day in the classroom, beginning of the day on the public address system, sporting events, baccalaureate services, and graduation ceremonies.

Table 9
Comparison by Selection on School-Sponsored Prayer Items

Item	*n*	Supt. % support	Bd. Member % support	χ^2	*p*
Beginning the school (classroom prayer)	205	41.1	71.0	24.501	<.001*
Beginning the school day (prayer on public address)	198	33.3	55.6	23.048	<.001*
Moment of silence (school-wide or by class)	202	60.0	75.0	9.692	.046
Sporting events	208	56.3	82.3	25.621	<.001*
Baccalaureate service	208	67.0	88.5	14.656	.005*
Graduation ceremony	208	61.6	86.5	19.428	.001*
Prayer clubs	207	62.5	71.6	10.405	.034*

$p < .05$, adjusted alpha= .0071

Question 5 was: How do respondents rate the influence of U.S. Supreme Court decisions, laws passed by the U.S. Congress and the Alabama Legislature, and community sentiment on their decision-making in the area of school prayer, and are these different between positions, and are there differences between methods of selection? Table 10 presents the percent for each response category for the four items. All four items had more than 75% of the respondents indicating the influence was important. U.S. Supreme Court decisions was highest (83.3% in the important range), followed by laws passed by the U.S. Congress (79.3% in the important range), community sentiment (78.8% in the important range), and laws passed by the Alabama Legislature (75.8% in the important range). There were no significant differences on these items between superintendents and board members nor between appointed and elected school officials.

Table 10
Percent Response on Influence Items, Total Group

		Not Important		Neutral	Very Important	
Item	*n*	1	2	3	4	5
U. S. Supreme Court decisions	209	6.7	1.0	9.1	14.8	68.4
Laws passed by U. S. Congress	208	6.7	2.4	11.5	19.7	59.6
Laws passed by Alabama Legislature	207	8.7	2.4	13.0	24.2	51.7
Community sentiment	208	4.8	2.9	13.5	22.1	56.7

Discussion

Support among school decision-makers for school prayer is generally high. Support is higher for student-initiated prayer activities as compared with school-sponsored prayer. This is consistent with court and U.S. Department of Education directives. Placing the origin of religious related activities on students rather than being school sanctioned is considered less objectionable to many public school stakeholders and less fraught with legal perils. It is also true that this approach is harder to attack. While superintendents and board members are held accountable for school policies and activities and as such would be the focus of praise or criticism from stakeholders, students have no organization that can be held accountable. Student-initiated activities that are reasonably considered within the realm of acceptable practice are rarely challenged.

Superintendents had higher differences than board members in support for student-initiated prayer as compared with school-sponsored prayer. Board members had high support for both student-initiated and school-supported prayer activities. Superintendents were high on support for student-initiated prayer but were lower on support for school-sponsored prayer. Also, relative to school-sponsored prayer, board members were more supportive than superintendents. Clearly, superintendents are more concerned about prayer activities being sanctioned by the school as opposed to being student-initiated. Another factor probably relates to the structure of school governance. Superintendents stand alone while board members stand as a group. In controversial issues, the superintendent is more cognizant of the "lone ranger" status that comes with the job and as such may tend to be more conservative on these issues and may be very cautious about taking stands on issues with legal ramifications. Board members may be more willing to take less conservative positions, because their opinions can be mediated by the group.

Relative to student-initiated prayer and school-sponsored prayer, elected superintendents and board members are more supportive than appointed superintendents and board members. Elected school officials must be more aware of the sentiments of those who elect them. As such, particularly in southern "Bible belt" states, they believe they will please more voters by being in support of school prayer than being against it. They may not know the numbers represented by the *Phi Delta Kappan* Gallup Poll (Gallup & Clark, 1987), but they do know there is generally support in their communities for school prayer activities. Of course, it needs to be kept in mind that these results have limited generalizability. Such attitudes may be very different in different regions of the nation. They are, however, likely to be very similar to attitudes in other southern states.

The perceptions of superintendents and school board members on any issue can have an impact on the students and the parents they serve. If those perceptions differ, whether it is because they are elected or appointed or because they perform different roles in the process of policy determination or decision-making, those differences have the potential for creating problems for all of those concerned. Although the laws pertaining to public school prayer differ from Supreme Court decisions, to laws passed by Congress, to individual state laws, it is still believed by most legal scholars that school-sponsored prayer is illegal (Zirkel, 1984). Typically, superintendents are more versed on legal issues relating to schools; however, many school board members attend training sessions provided by state associations designed to inform them about current issues. Nevertheless, it could be argued that school board members, particularly elected ones, are going to take positions on issues that are more in tune with their constituencies than will superintendents.

The results of this indicate that school board members have higher support for school-sponsored prayer, advocating a practice that has been determined to be illegal. It is prudent for superintendents to be sensitive to the will of the school board and the community while still attempting to uphold the law, whether they personally agree or disagree with the law. Superintendents who have been unsuccessful at accomplishing this typically find themselves seeking other employment opportunities. If more school boards become elected rather than being appointed, while a high majority of superintendents are appointed, there will be more potential for conflict over this issue. One method for ameliorating this situation is for superintendents to promote student-oriented prayer as an alternative and legal substitute for school-sponsored prayer. Another recommended practice is to facilitate the training of school board members in the legal aspects and successful practices used by other school systems in dealing with this issue. Clearly, compromise and educational opportunities offer the best hope to superintendents and board members in making decisions relating to public school prayer which uphold current law and protect the First Amendment rights of public school students.

References

Alley, R. S. (Ed.). (1988). *The Supreme Court on church and state.* New York: Oxford University Press.

Barber, P. E. (1992). Bishop v. Aronov: "No talking in class!": Does the elementary school adage apply to university professors? *Alabama Law Review, 44,* 211-235.

Brown, C. S. (1983). No amen for school prayer. *Learning, 12*(1), 42-43.

Cubberley, C. P. (1962). *Public education in the United States.* Cambridge, MA: Houghton-Mifflin.

Douglas, W. O. (1966). *The Bible and the schools.* Boston: Little Brown.

Gallup, A. M., & Clark, D. L. (1987). The 19th annual Gallup poll of the public's attitude toward the public schools. *Phi Delta Kappan, 69,* 17-30.

Horner, J., & Barlow, B. (1994). Prayer in the public schools in light of Lee v. Weisman. *West's Education Law Reporter, 87,* 268-273.

Kelley, D. M. (1984). School prayer and true religious liberty. *The Education Digest, 50,* 52-55.

LaNoue, G. R. (1967). The conditions of public school neutrality. In T. R. Sizer (Ed.), *Religion and public education* (pp. 22-36). Boston: Houghton Mifflin.

Lowry, C. W. (1963). *To pray or not to pray.* Washington, DC: The University Press.

Nasaw, D. (1979). *Schooled to order: A social history of public schooling in the United States.* New York: Oxford University Press.

O'Reilly, R. C., & Green, E. T. (1992). *School law for the 1990's: A handbook.* New York: Greenwood Press.

Pfeffer, L. (1975). *God, Caesar, and the Constitution.* Boston: Beacon.

Rankin, N. R., & Strope, J. L. (1994). Prayer at public school graduation: What is a school official to do? *West's Education Law Reporter, 3,* 464-467.

Schamel, W. B., & Mueller, J. W. (1989). Abington v. Schempp: A study in the establishment clause. *Social Education, 53,* 61-66.

Sendor, B. (1983). Prayer gets a foot in the schoolhouse door. *American School Board Journal, 170*(11), 25.

Starr, I. (1985). My pilgrimage to the wall of separation. *Update on Law-Related Education, 9*(2), 2-5; 37-48.

Thomas, W. L., & Anderson, R. J. (1982). *Sociology: The study of human relationships.* New York: Harcourt Brace.

U. S. Department of Education. (1995). (Untitled letter from Richard W. Riley to all public school superintendents and school officials.) Washington, DC.

Vacca, R. S., & Hudgins, H. C., Jr. (1994). Pomp and controversy. *American School Board Journal, 181*(5), 29-32.

Zirkel, P. A. (1984). A hypothetical case in the public school. *The Clearinghouse, 57,* 346-347.

Zirkel, P. A. (1994). Graduation invocations and benedictions: Good faith interpretations? *West's Education Law Quarterly, 3,* 472-484.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 17-34

Written Expression Reviewed

Jason C. Cole
California School of Professional Psychology, San Diego, California

Kathleen A. Haley
Boston College

Tracy A. Muenz
California School of Professional Psychology, San Diego, California

Recent research by Cole, Muenz, Ouchi, Kaufman and Kaufman (1997) on a theoretical model of written expression by Hooper et al. (1994) has led to the suggestion of specific criteria for measures of written expression. In this review, 12 recently published measures of written expression were assessed for their ability to meet criteria detailed by Cole et al.: specific stimulus criteria presented by Hooper et al.; the discriminating ability of the scoring criteria; direct modality of assessment; and desirable psychometric properties of the instrument. The types of stimuli used in the 12 measures were found to vary from meeting none of Hooper et al.'s criteria to meeting a few; none of the measures met all stimulus criteria. Only one of the scoring systems was found to fully provide direct interpretation of the differences between expert and poor writing quality; two other measures partially allowed for this interpretation. Additionally, 9 of the 12 measures contained the more valid direct assessment; three of these nine contained both direct and indirect assessment modalities. Most of the reported reliabilities for these measures were either improperly assessed or too low for psychodiagnostic utility. Ultimately, none of the assessment tools reviewed met all of the evaluative criteria.

Written expression assessment has been a vagary that has eluded proper analysis for over a decade. Much of the problem has been due to a lack of theoretical postulation regarding written expression. Recently, more refutable theories have been presented in the written expression literature. Hooper et al. (1994) claimed that the lack of a theoretical model has led to many well-intentioned, but poorly designed, written expression assessment tools. One of the suggestions made by Hooper et al. is that the stimulus used in the elicitation of a written response should contain specific criteria. Namely, the stimulus should be a photograph, portray at least two characters (preferably a protagonist and possibly an antagonist), display a novel and interesting depiction, and portray a state of conflict such that a subsequent goal-directed sequence of events is necessary in order to resolve the conflict. These criteria should elicit more thematic, goal-directed responses than traditionally used stimuli (e.g., line drawings or auditory prompts).

Cole, Muenz, Ouchi, Kaufman, and Kaufman (1997) tested Hooper et al.'s hypotheses and found strong support. In Cole et al.'s study, 50 subjects from Southern California were each administered two visual stimuli for the elicitation of a written response. The first stimulus was the "Box" prompt, a line drawing from the Peabody Individual Achievement Test - Revised. It depicts a man delivering a wooden box to a house; the door is open, a dog and cat scour near the threshold, and two children approach from the distance. A second prompt was developed by the authors to meet the stimulus criteria proposed by Hooper et al. (1994). This color photograph, called the "Cliff," depicted two men on a set of rocky cliffs; one is in need of help and reaching to the other while two onlookers watch from the beach below. Instructions from the PIAT-R manual Level II Written Expression Subtest were generally followed. Items (scoring criteria) were grouped into two categories. The first set of items, called "Structure" items, assessed the quality of writing for cohesiveness, organization, and development of ideas. The second set of items –

Jason C. Cole, Department of Clinical Psychology, California School of Professional Psychology, San Diego, California; Kathleen A. Haley, Department of Educational Research, Measurement and Evaluation, Boston College; Tracy A. Muenz, Department of Clinical Psychology, California School of Professional Psychology, San Diego, California. The authors sincerely thank Alan S. and Nadeen L. Kaufman, and Bryan Y. Ouchi. Also, special thanks to Tom R. Smith and Nusheen Cole. Correspondence concerning this article should be addressed to Jason C. Cole, PO Box 2664, Laguna Hills, California, 92654-2664. Electronic mail may be sent via Internet to JasonCCole@home.net.

"Mechanics" items – assessed grammar, punctuation, and writing legibility. Cole et al. hypothesized that Structure items, according to Hooper et al's (1994) hypothesis, would receive higher ratings on the "Cliff" stimulus than on the "Box." However, the "Mechanics" items were hypothesized to show no difference between the stimuli. Both hypotheses were supported.

These results found by Cole et al. (1997) regarding the strength of Hooper et al.'s (1994) stimulus criteria were then combined with criteria drawn from other theories of written expression. Cole et al. postulated that the combination of four criteria would lead to better instruments in written expression assessment: the use of a proper stimulus (as described above), the ability of the scoring criteria to differentiate among writing qualities, the use of direct assessment, and the proper establishment of sound psychometrics.

One of the aforementioned criteria for written expression was a combination of many studies on the qualities of "expert" and "poor" writers (see Cole et al., 1997). Previous researchers found that the writing characteristics of those persons historically considered to be experts were likely to contain the following: goal directed writing, fluid transitions, an understanding of the writing assignment, and organizational skills (Bereiter, 1980; Burtis, Bereiter, Scardamalia, & Tetroe, 1983; Fitzgerald, 1987; Flower & Hayes, 1981; Halliday & Hasan, 1976; Hayes & Flower, 1986; Hooper et al., 1994; McCutchen & Perfetti, 1982; Scardamalia & Bereiter, 1986; Sommers, 1980). Furthermore, other studies have shown specific characteristics recurrent in the writing of those considered to be poor writers: poorly organized text at both the sentence and paragraph level; decreased likelihood to modify spelling, grammar or the substance of their writing in order to enhance communication of their ideas (and thus poorer communication); and stories that were less likely to be interesting (Anderson, 1989; Englert, 1990; Englert, Raphael, Anderson, Gregg, & Anthony, 1989; Englert & Thomas, 1987; Graham & Harris, 1987; Graham & Harris, 1989; Graham, Harris, & MacArthur, 1990; Graham, Harris, MacArthur, & Schwartz, 1991; Hooper et al., 1994; Wong, Wong, & Blakenship, 1989). Cole et al. (1997) advocated the use of maximal differentiation of the "expert" and "poor" writing qualities in order to increase the validity of written expression assessment. This differentiation will increase the construct validity of a measure by attempting to measure traits crucial to writing quality.

The type of assessment has also been postulated to have an impact on the viability of written expression. Breland and Gaynor (1979) and Auchter and Hatch (1990) have suggested that the overall validity of indirect assessment (where responses are generally limited to a sentence, or two, at the most) was too low to be effective in measures of written expression. Therefore, the inclusion of direct assessment (which allows responses in multi-paragraph form[1]) of written expression should also be a critical aspect to a written expression assessment tool.

Finally, psychometric properties should be strong (see Anastasi, 1988), despite the difficulty of achieving high reliability with direct assessment (Auchter & Hatch, 1990; Breland & Gaynor, 1979). Many authors of written expression measures report coefficients of impressive magnitude for their psychometric properties. However, closer inspection reveals some serious methodological flaws. Therefore, not only should the psychometric properties be sound, the psychometric analyses should be conducted with sound methodology.

Cole et al. (1997) advocated the use of the aforementioned criteria in the creation of written expression batteries. However, these criteria of written expression have not been employed as a complete system for the development of any currently available measure of written expression. Although many of these theories are new, and some of the empirical validation is even newer, current measures of written expression are likely to contain aspects of the aforementioned theories and criteria. This review rates 12 recently published measures of written expression on the aforementioned criteria, specifically: (1) Hooper et al.'s (1994) stimulus criteria; (2) a means for differentiation of "expert" and "poor" writing qualities, specifically in the scoring system of the measures; (3) the use of direct assessment of written expression; and (4) reasonable psychometric properties and proper assessment of these properties.

Hooper et al.'s (1994) stimulus criteria will be rated for each specific recommendation. That is to say, the 12 measures will be reviewed for their (1) type of stimulus (e.g., the recommended color photograph versus another type of stimulus), (2) portrayal of at least two characters, (3) depiction of a novel and interesting scene,[2] and (4) portrayal of a state of conflict such that a subsequent, goal-directed sequence of events is necessary in order to resolve the conflict (see Footnote 2). The discriminability of the scoring criteria was categorized as sufficient if the scoring system had a readily interpretable means for comparing "expert" and "poor" writing qualities. Therefore, it wasn't adequate for the scoring system to simply have items which differentiated the "expert" and "poor" writing abilities – the system must have also provided some normative scale assessing these abilities. Measures were examined for their modality of assessment – whether a response was direct or indirect – by the amount of writing required. To qualify as a direct form of written expression the measure must allow multi-paragraph responses (see Footnote 1). Last, reliability coefficients should exceed .80 (see Anastasi, 1988), and all psychometric properties should be assessed with sound methodology. The efficacy of the psychometric assessments will be addressed on a test by test basis.

Assessment of Written Expression Measures

The PIAT-R

The Level II Written Expression subtest of the Peabody Individual Achievement Test - Revised (PIAT-R; Markwardt, 1989) provides a popular measure of written expression in a direct measurement format. This measure was also instrumental in the Cole et al. (1997) study in order to affirm Hooper et al.'s (1994) hypotheses regarding necessary stimulus criteria.

Stimulus criteria. There are two prompts labeled "A" and "B" which facilitate alternate form administration. One stimulus contains a line drawing of a delivery man wheeling a large wooden crate up to a doorstep, and the other stimulus is a drawing showing several people responding to wind-blown money from a woman's dropped purse. Neither stimulus meets Hooper et al.'s (1994) first criteria as they are both line drawings. Both prompts do contain more than one character. However, there is no obvious protagonist or antagonist in either prompt. Both stimuli could be considered interesting to the examinee, but a woman losing her purse or a delivery man making a delivery are hardly novel. Finally, only the stimulus containing the money in the street meets Hooper et al.'s last criteria, as the woman who has dropped her purse may need to initiate a goal based sequence of events to reacquire the purse. The prompt containing the delivery man does not appear to have a potential conflict situation which would compel a character to initiate a goal based sequence of actions.

Discriminating ability of the scoring criteria. The PIAT-R contains items which assess both "expert" and "poor" writing qualities. However, many items are also present in the PIAT-R scoring system that do not differentiate "expert" and "poor" writers. Therefore, the PIAT-R does not provide a normative scale that can provide information on the differentiation of "expert" and "poor" writing qualities.

Modality of assessment and psychometric properties. The PIAT-R also suffers from a typical problem with direct assessment of written expression -- its reliability is low. In fact, the interrater reliability range of the PIAT-R (median $r = .58$ for the "Money in the Street" prompt, median $r = .67$ for "The Box" prompt) (Markwardt, 1989, p. 71) is not a very large effect size, according to Cohen (1990; 1992). A very large effect size is needed to explain at least 50% of the variance, while only 33.6% and 44.9% of the variance is explained in the "Money in the Street" and "The Box", respectively. Making decisions about educational placement for learning disabled children with a measure having low reliability, and thus low percentage of explained variance, is a questionable practice. Although the internal consistency reliability is higher (mean coefficient alphas were .86 and .88), Kaufman (1990, p. 627) noted that the use of coefficient alpha was not appropriate with the PIAT-R as items were not experimentally different. In other words, the use of coefficient alpha necessitates that items scored are not dependent, in any known manner, on other items. Yet, in the Written Expression subtest of the PIAT-R all scoring items are based on the same stimulus, and thus a known dependency exists. Further, in learning disability assessments, a reliability measure sensitive across individuals is more useful than one that shows consistency for a single individual. Thus, the internal consistency reported for PIAT-R is of little benefit.

Content and construct validity are both examined in the PIAT-R manual. Content validity was assessed by expert judges used throughout the developmental process of the PIAT-R. Markwardt (1989) also found the content validity to be sound via the internal consistency assessments (split-half and Kuder-Richardson). However, as discussed above, this use of internal consistency was deemed to be inappropriate for the written expression subtest. Construct validity of the PIAT-R as a whole was assessed with developmental changes, convergent and factor analytic validity (see Anastasi, 1988; Benes, 1992; Rogers, 1992). However, the Written Expression subtest was not included in any of the quantitative validity studies. According to Markwardt (1989, p. 52) the omission of the Written Expression subtest from the validation process was due to restricted range, and an achievement growth curve which showed less growth at older grades than younger grades.

Conclusion for the PIAT-R. The PIAT-R contains a direct method of assessment where the line drawings meet a few of Hooper et al's criteria. However, the PIAT-R's inability to differentiate between "expert" and "poor" writers and overall poor psychometric qualities limit its usefulness as a test of written expression.

The WIAT

The Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992) is another achievement battery that contains a written expression subtest. The

WIAT has been touted by its publisher as the only achievement test standardized with the Wechsler Intelligence Scale for Children - Third Edition (WISC-III; Psychological Corporation, 1991). Thus, the popular WISC-III has a companion: the WIAT.

Stimulus criteria. The WIAT was one of only four measures assessed in this article that did not provide a pictorial stimulus. Instead, the writers are encouraged to create their own mental image of a scene and write about what they envision. One prompt asks subjects to write about their dream house and the second prompt asks subjects to write a letter to a friend asking them to come along on a trip the subject has wanted to take.

Stimuli for the WIAT do not meet any of this article's stimulus criteria. The stimuli are not photographic, do not contain two or more characters, do not depict an interesting and novel situation, nor do they present a state of conflict. Moreover, another problem exists with the stimuli from the WIAT. Visualization should not be a necessary component of adequate writing quality. Yet, visualization appears to be a necessary component for a proper response to the stimuli from the WIAT. One has to create a visual image of what they want to write about, and then write it. Most other written expression batteries only ask that subjects write about what they see in the stimuli – they needn't visualize the stimuli first.

Discriminating ability of the scoring criteria. Although items exist in the WIAT scoring systems that have "expert" and "poor" writing differentiation ability, the scales were not developed to be used in this fashion. The Analytic Scoring system contained other items which did not differentiate the aforementioned qualities. Further, the Holistic Scoring system did not contain criteria which accurately differentiated writers. Thus, a normative scale differentiating "expert" and "poor" writing qualities was not provided in the WIAT.

Modality of assessment and psychometric properties. The WIAT, which was designed to allow multi-paragraph responses, provides a direct assessment of written expression. Hence, the WIAT met the criteria for modality of assessment.

The reliabilities given in the manual for the WIAT are suspiciously high. Internal consistency measures are reported for the WIAT in its manual; however, these findings suffer from the same problems as the PIAT-R regarding Kaufman's (1990) critique of the use of internal consistency assessment. The stability of the WIAT has an average (using Fisher's z transformations) $r = .77$, which is corrected for range restriction. Interrater reliability results from the WIAT manual have an average $r = .79$ and .89, within the different prompts. Although these results appear to be adequate reliability estimates (Anastasi, 1988), and high for a direct measure of written expression, they are spuriously inflated due to the large age range used in the interrater reliability computation (kindergarten through 12th grade). Inflation, or a spuriously high correlation, occurs when an exceedingly wide range of scores is correlated, especially when the ability in question (in this case, written expression) is subject to substantial growth across the age span. Therefore, while the reported reliability of the WIAT is within an acceptable range according to the standards of both Cohen (1992) and Anastasi (1988), the methodology for estimation is not adequate.

Validity of the Written Expression subtest for the WIAT was demonstrated with convergent, concurrent and discriminative validity. Convergent validity showed a correlation of .72 between the WIAT and the Woodcock-Johnson Psychoeducational Battery - Revised (WJ-R/A; Woodcock & Johnson, 1989, see review below). The sample contained ages from 7 to 14, with a median age of 12, and 77% of the sample was male. The sample was over-representative for males and has a large heterogeneous age group that is likely to have spuriously inflated the correlation. Criterion-related validity, specifically concurrent validity, was assessed by correlating subjects' scores on the WIAT to the subjects' academic grades. These correlations were low: reading $r = .34$, math $r = .34$, and language $r = .36$. Discriminative validity was based on scores from gifted, mentally retarded, emotionally disturbed, learning disabled, Attention Deficit Hyperactivity Disorder, and hearing-impaired children. The mean score for gifted children was 114.5, whereas mentally retarded children had a mean of 77.6. Emotionally disturbed and learning disabled children had mean scores of 83.0 and 85.1, respectively. Children with Attention Deficit Hyperactivity Disorder and hearing impairments scored 92.8 and 90.6, respectively. Hence, discriminative validity is strong with the WIAT Written Expression subtest.

Conclusions for the WIAT. The WIAT contains scoring items that allow for some differentiation between "expert" and "poor" writing qualities, and is a direct measure of written expression. The scale, however, was not normed to be used for the differentiation of the aforementioned writing qualities. Because of the use of an auditory prompt, questionable reliability estimates and the problems with the scoring system, this measure only meets one of the four criteria proposed by Cole et al. (1997).

The WJ-R/A

The Woodcock-Johnson Psychoeducational Battery - Revised/Achievement (Woodcock & Johnson, 1989) is another very popular achievement battery (Harrison, Kaufman, Hickman, & Kaufman, 1988) that contains a measure of written expression. The subtest Writing Samples is an indirect measure of written expression in which the subject writes a brief response to an array of stimuli in order to demonstrate writing ability (Kaufman, 1990).

Stimulus criteria. Many stimuli are presented in the WJ-R/A. The stimuli are presented either as text or as a line-drawing, combined with auditory instructions. An example of the text stimuli is an item that asks for the steps a father would follow when building a house. Visual stimuli are comprised of small black-and-white line-drawings, such as a baby bird coming out of an egg, or two children playing catch.

The combination of textual and auditory stimuli presentation from the WJ-R/A inhibited these stimuli from meeting the stimuli criteria. Line-drawings contained in the WJ-R/A did not consistently meet any of the stimulus criteria. In fact, all of the items failed to meet the criteria for type of stimulus and a depiction of a novel and interesting scene. Some of the stimuli did present two or more characters, and some of the stimuli did portray a state of conflict. However, these last two criteria were not consistently portrayed across all visual prompts. Therefore, none of stimulus criteria of Hooper et al. (1994) were met by the WJ-R/A.

Discriminating ability of the scoring criteria. Scoring of the written responses in Writing Samples is conducted in a holistic type approach. An overall score is given to each brief response, and each score is derived by comparing it to a set of templates provided in the manual. The holistic scoring criteria are somewhat vague. Whereas the supplement to Writing Samples by Mather (1993) established more basic criteria for the scoring of the written responses, the level of detail left some ambiguity. Specific points are sought out on each item, but how to handle additional improvements or impairments in the response was unclear.

Use of different specifications for each item, and an ambiguous holistic approach for scoring the written responses influenced the authors of this article to rate this test as poor for differentiation. The standardized scores for Writing Samples do not provide information on the differentiation of "expert" and "poor" writing qualities. Moreover, none of the items contained in the scoring specifically tap this critical criterion proposed by Cole et al. (1997).

Modality of assessment and psychometric properties. The WJ-R/A presents an indirect assessment of written expression. Although more than a single sentence can be written in the space provided for the response, scoring of the items was limited to the most appropriate sentence in the response (Mather, 1993).

The WJ-R/A does have high reliability: median r = .89 (Woodcock & Mather, 1989, Table 6-5). Breland and Gaynor (1979) note that not only do indirect measures of written expression usually yield higher reliabilities in written expression assessment, but they also have high face validity and credibility with the English teaching community. Yet, construct validity limitations (Anderson, 1989; Auchter & Hatch, 1990) of indirect measures of written expression are such that direct measures are preferable.

Conclusions for the WJ-R/A. Overall, while the WJ-R/A offers a psychometrically sound measure of written expression, this is the only criterion met of those proposed by Cole et al. (1997). The stimuli did not meet any of the stimuli criteria, differentiation of "expert" and "poor" writing qualities was not possible, and the WJ-R/A provided an indirect modality of written expression assessment.

The TOWL-2 and TOWL-3

The Test of Written Language – Second Edition (Hammill & Larsen, 1988) is another common assessment tool in the written expression assessment arsenal. After several changes to the TOWL (Hammill & Larsen, 1983) were made, the new version received acclaim for the authors' results (Benton, 1992). Various components of writing are assessed in the TOWL-2, including conceptual, conventional, and linguistic components. Further, the assessment is conducted in both formed and spontaneous formats that are also direct and indirect assessments, respectively.

The Test of Written Language – Third Edition (TOWL-3; Hammill & Larsen, 1996) was released by the test's publisher during the creation of this article. A major emphasis during the creation of the TOWL-3 was the limitation of test bias. A review of the TOWL-3 is provided concurrently with the TOWL-2 review as examiners seeking information on the TOWL-2 may have not yet converted to the newest revision.

Stimulus criteria. The TOWL-2 and TOWL-3 use the same stimuli. The tests provide two black-and-white line drawings for pictorial stimuli in the spontaneous (direct assessment) section of the test. The first drawing depicts a lunar scene in which an astronaut in the forefront of the drawing is carefully examining something on the ground; several colleagues are approaching from the distance, and much space activity is occurring in the periphery of the drawing. The second drawing depicts a prehistoric battle where several humans attack a herd of mastodons. Peers of the attackers are waiting in the distance while three main members attack the herd.

These drawings compare well with some of the criteria set forth by Hooper et al. (1994). Although the stimuli are not color photographs, they both contain at least two characters and have novel and interesting depictions. The prehistoric battle scene does present a state of conflict. A state of conflict may be present in the lunar scene – this interpretation is certainly subjective (see Footnote 2). The scene, however, appears to instill a need for goal-directed activity. Thus, this criterion was consistently met for both stimuli. Ultimately, the TOWL-2 and TOWL-3 stimuli met three of four stimuli criteria.

Discriminating ability of the scoring criteria. The TOWL-2 does not contain a normed scale that differentiates "expert" from "poor" writers. Some items used to evaluate the writing of the spontaneous task may be used to differentiate the "expert" and "poor" writing qualities of subjects. However, the inclusion of other items in this scale precludes it from being used as a normative scale to differentiate "expert" from "poor" writers.

The TOWL-3, however, was the best test reviewed in this article for its ability to differentiate writing qualities. In fact, it was the only measure that fully met this criterion. The Spontaneous Subtests contain three sections, Contextual Conversions, Contextual Language, and Story Construction. The Contextual Conversions section measures items relating to grammar and punctuation. Contextual Language measures a subject's use of language in writing. Finally, Story Construction evaluates story fluidity and thematic presentation.

Story Construction scores present a normative evaluation differentiating "expert" and "poor" writing qualities. This scale evaluates all of the writing qualities found in "expert" writers, and also assesses qualities commonly found in poor writers. Overall, the content of the Story Construction scale matched this article's criterion for discriminability amongst writers very well.

Modality of assessment and psychometric properties. Direct modalities of written expression assessment are provided in the TOWL-2 and TOWL-3. Moreover, each of these tests provides indirect assessment. A combination approach of the two modalities may provide beneficial information, depending on the circumstances of one's evaluation.

The reliability of the TOWL-2 has been called both "sound" (Benton, 1992) and "inflated" (Ryan, 1992). Interrater reliability ranged from .91 to .99 between 2 raters scoring 20 protocols. Generalizability is limited enough by having only two raters; moreover a conflict may exist in that one of these raters was an author of the test. Stability and consistency were found to be adequate for the Overall Written Language quotient, but these calculations were made with excessive heterogeneity that leads to inflation (Ryan, 1992). Overall, the reliabilities of the TOWL-2 are questionable and overestimated.

The validity estimates of the TOWL-2 were even more precariously assessed than the reliability. Neither the Spontaneous Writing Quotient nor any of the spontaneous subtests correlate with scores on the SRA Achievement Series Language Arts test above .49 (Benton, 1992). The construct validity of the TOWL-2 showed moderate inter-subtest correlations while varimax rotation on factor analysis showed the Spontaneous and Contrived Writing factors. However, Ryan (1992) challenged the findings of the unrotated factor analysis by noting that only one factor appears to emerge instead of the three factors hypothesized by the authors. Hence, a rotation to the resulting construct presented by Hammill and Larsen (1988) may have been inappropriate.

The authors of TOWL-3 assessed the test's internal consistency with coefficient alpha. This method was previously noted (see the PIAT-R and WIAT sections) as being inappropriately used when only a single stimulus is presented for which many scoring items were scored. Equivalency of the two forms (A and B) of the TOWL-3 was assessed with an alternate forms assessment correlation. Correlations for the forms across subtests and ages ranged from .72 (7 year-olds on Sentence Combining), to .94 (17 year-olds on Story Construction), with almost all correlations falling in the acceptable range of .80 or higher. However, the above range does not include Contextual Conventions, which has a markedly poorer form equivalency: *r*s ranged from .60 to .75, too low according to the criteria of Anastasi (1988). Caution should be taken by examiners who need to compare test administrations across the forms for the Contextual Language subtest. Test-retest reliability for TOWL-3 was assessed with less than admirable methodology. Two grades were selected from the Austin, Texas area over a two week interval. No reason for using these grades was provided, other than that they were *different*. Moreover, the generalizability was poor for two reasons. Only one geographic area of the nation was sampled, and only 2% (approximately) of the normative sample used in the test-retest analyses. Interrater reliability of the TOWL-3 was poorly assessed. Two PRO-ED[3] staff, hardly generalizable to the typical rater, rated 38 protocols. The percentage of protocols – approximately 3% – drawn from their normative sample for the interrater reliability analysis was, again, too low. Correlations between the scores of raters were not presented with the means and standard deviations, nor were the analyses assessed with an intraclass correlation (see Shrout & Fleiss, 1979). Therefore, while the correlations amongst the two raters ranged from .80 (Form A for Story Construction) to .97 (many subtests), a difference in means between the examiners may have existed. Further, the correlations were spuriously inflated as all grade levels were combined into a single correlation.

Content Validity of the TOWL-3 was assessed with item bias procedures in order to detect items biased against minorities. Differential Item Functioning (DIF) was assessed by Hammill and Larsen using a delta score approach by Jensen (1980). Results of the analysis were very high (see MacEachron, 1982). Whereas the delta scores approach is useful, and Hammill and Larsen should not be criticized for using this method, more efficient methods are available. According to Osterlind (1983) an Item Response Theory approach is highly useful in the detection of DIF. More powerful analyses were available for the authors of the TOWL-3 (see Holland & Thayer, 1986; Holland & Thayer, 1988; Lord, 1977a; Lord, 1977b; Stout & Roussos, 1995). Furthermore, a key assumption in the assessment of DIF is unidimensionality (Osterlind, 1983). Yet Hammill and Larsen provide many standardized scales of written expression assessment in the TOWL3, indicative of many dimensions. Thus, a non-parametric test, with less constraint from the unidimensionality assumption in DIF, may have provided more accurate results for the TOWL-3's DIF analysis (Stout & Roussos, 1995). Concurrent validity for the TOWL-3 was assessed by comparing the TOWL-3 to the Comprehensive Scales of Student Abilities (Hammill & Hresko, 1994). This study assessed 76 students from a Texas elementary school. A detailed listing of the correlations was not provided by the authors. The correlations, however, were low. The median subtest correlation was .50, whereas the median composite correlation was .53. Ideally, these correlations should minimally be .60.

Summarily, the reliability and validity estimates for the TOWL-2 and TOWL-3 were precariously assessed. The many flaws in methodology result in unknown psychometric properties for both of these tests. Ultimately, these tests did not meet the criteria for sound methodology of psychometric assessment.

Conclusions for the TOWL-2 and TOWL-3. The TOWL-2 meets some of Cole et al.'s (1997) suggested criteria for measures of written expression – it has a direct assessment component, a visual stimulus meeting three of four stimulus criteria, and attempts to provide sound psychometric utility. Yet, the TOWL-2 does not allow appropriate differentiation of the "expert" and "poor" writing qualities of subjects, does not have a photographic stimulus, and has questionable psychometric properties. The TOWL-3 provided improvements. The scoring system does allow for direct interpretation of the differences between "expert" and "poor" writers, the stimuli met three of four criteria, and the test offers a direct modality of assessment. However, proper psychometric analyses still need to be conducted for a more refined estimation of the TOWL-3's reliability and validity.

The OWLS

The Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1996) is a new assessment tool and contains a written expression section. The OWLS was developed with a special emphasis on minority sensitivity and fairness, and the items were comprehensively assessed both quantitatively and qualitatively.

Stimulus criteria. Stimuli for the OWLS consist of text read aloud by the examiner. The content of the auditory stimuli ranges from a request to write a brief letter to one's mother regarding breaking something in the house, to filling in missing pieces of sentences. None of the stimulus criteria are consistently met across the variety of prompts provided in the OWLS.

Discriminating ability of the scoring criteria. The Content Category has marginal ability for differentiation of the "expert" and "poor" writing qualities of subjects. The low frequency of each quality differentiating "expert" from "poor" writers and the marginal overlap between the Content Category and the qualities differentiating "expert" from "poor" writers, suggest that this scale will not adequately differentiate. Primarily, problems exist in using this scale as a differentiating scale for the "expert" and "poor" writing abilities due to a lowered reliability (via decreased item total in the scale) and other items assessed in this scale which do not differentiate the writing qualities.

Modality of assessment and psychometric properties. The OWLS offers several lengths of subject responses, yet none of the responses exceed a paragraph in length. Thus, the OWLS is more appropriately labeled an indirect measure of written expression rather than a direct measure (see Footnote 1).

The psychometric properties of the OWLS are consistent with indirect written expression assessment. The internal consistency has a mean $r = .87$ with a range from .77 (for the 19-21 year old subject range) to .94 (for the 7 year old subjects). Stability measures for the OWLS were assessed over a range of 8 to 165 days, with a 9 week median. Corrected coefficients (using a formula by Guilford, 1954, p. 392) for two age groups were .88 (for ages 8-0 to 10-11) and .87 (for ages 16-0 to 18-11). The stability estimates do contain questionable sample composition. The reason for restricting the age ranges, and then dichotomizing them was not explained in the manual. Furthermore, 61% of the subjects were from the South. Interrater reliability of the OWLS was assessed using four raters with no previous experience rating written expression responses. Although the raters were given training similar to a typical examiner, the assessment had other weaknesses. Each age cohort consisted of only 15 subjects (from 3% - 7% of each age cohort was assessed) – too few for stable reliability coefficients. Furthermore, as written expression assessment is plagued with interrater reliability difficulties, a much higher percentage is admirable (for example, see Ouchi, Cole, Muenz, Kaufman, & Kaufman, 1996). Therefore, while the interrater reliability intraclass correlations (see Shrout & Fleiss, 1979) ranging from .91 (for ages 12 to 14) to .98 (for ages 5 to 7) were admirable, the procedures used to assess interrater reliability were haphazard.

Content validity was assessed by authorities from many minority groups. Concurrent validity was evaluated using the Kaufman Test of Educational Achievement (K-TEA; Kaufman & Kaufman, 1985), the Peabody Individual Achievement Test - Revised (PIAT-R; Markwardt, 1989) the Woodcock Reading Mastery Test - Revised (WRMT-R; Woodcock, 1987), the Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981), and the Clinical Evaluation of Language Fundamentals Revised (CELF-R; Semel, Wiig, & Secord, 1987). Written Expression on the OWLS correlated (correcting for range restriction) well with both the K-TEA and the WRMT-R (all subtests correlated in the .80s). The PIAT-R correlations ranged from .63 (General Information) to .84 (Total Test), and the PIAT-R had lower overall correlations than the previous two tests. The Written Language Composite in the PIAT-R correlated with the OWLS at .71, showing fair comparability between the two tests. The PPVT-R and the OWLS correlated at .62, while the range of correlations for the CELF-R ranged from .55 (Oral Directions) to .85 (Sentence Assembly). Discriminant validity was assessed with the Wechsler Intelligence Scale for Children - Third Edition (WISC-III; Psychological Corporation, 1991) and the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990). Written Expression correlated to the WISC-III at .72, .64, and .70 for Verbal IQ, Performance IQ, and Full Scale IQ, respectively. Similarly, Written Expression correlated to the K-BIT at .67, .41, and .58 for Vocabulary, Matrices, and the K-BIT Composite, respectively. Discriminant validity data are shown in Table 9.9 of the OWLS manual (Carrow-Woolfolk, 1996). Ultimately, the discriminant and concurrent validities were equally as high, as the discriminant validity coefficients were higher than is desirable. The OWLS is shown to have significant ability to discern a language-impaired group from a normal group. Unfortunately, the methods were reported in the manual for the analysis of construct validity. Although the validity assessments in the OWLS manual may be somewhat overstated due to heterogeneity of groups, the extensive work on validation is commendable and psychometric properties are, on the whole, quite sound.

Conclusion for the OWLS. Although the OWLS offers reasonable psychometric properties for written expression assessment, this criterion is the only criterion Cole et al. (1997) suggested that is met by the OWLS. The OWLS uses indirect assessment with auditory prompts meeting none of the stimulus criteria. Furthermore, the scoring system can only be used to differentiate "expert" and "poor" writing qualities marginally, at best.

The WLA

The Written Language Assessment is a direct test of written language created by Grill and Kirwin (1989). An excellent review of the WLA was written by Norris (1992).

Stimulus criteria. Three writing tasks are given to subjects; there is an expressive, a creative, and an instructive task. Each task has a different stimulus. The expressive task has a black-and-white photograph, the creative task has a picture (of a painting) and the third task contains written instructions to "Write how you would tell a little kid about the danger of fire." While this is the first test reviewed in this article that has a photo stimulus (even the picture of a painting comes closer to the criteria by Cole et al. (1997) than a line drawing), Norris notes that the stimuli are "not conducive to story telling" (p.677). Only the picture of a painting meets the criteria of being a color photo and having two characters. None of the aforementioned prompts were considered both "interesting and novel." Also, none of Grill and Kirwin's four prompts met the criterion for a state of conflict. Norris concurred by stating "The picture . . . does not suggest any problem or goal" (p.677).

Discriminating ability of the scoring criteria. As noted earlier the ability to generate "expert" quality writing is linked with the ability to generate stories. Scoring of the WLA responses is not helpful for differentiation of the "expert" and "poor" writing qualities of subjects. There are numerous scores in the WLA that are based on the number of words used in the composition. This technique is rather far from the criteria purported to be useful by Cole et al. (1997). For example, a writer may be verbose without understanding the meaning of the words used or the proper construction of a sentence.

Modality of assessment and psychometric properties. The WLA allows for multi-paragraphical responses and is therefore a direct assessment of written expression. Thus, the WLA met the modality of assessment criterion.

The internal consistency reliability of the WLA ranges from median coefficients of .61 (Readability) to .90 (Written Language Quotient - a composite scale). Again, internal consistency on items that come from the same stimulus is inappropriate. Furthermore, interrater reliability reaches 80% agreement (Moran, 1992). However, only 10% of the normative sample was selected for interrater reliability analysis; thus the WLA falls into the same problem of insufficient sampling as the OWLS does. While these reliabilities are high for direct assessments of written expression, they must be interpreted with an understanding of the scoring criteria. The scoring criteria for the WLA are based largely on word ratios, thus reliability should be high. However, the validity is concomitantly lowered as the *what* we are measuring becomes word ratios and not the ability to put these words together coherently.

Validity studies were not conducted by the authors. The authors of WLA stated that validity need not be assessed as the test only consists of writing (Grill & Kirwin, 1989, p.60). Cooper (1994) argued against any claim of demonstrated content validity simply because the content consists only of writing. Cooper stated that validity simply cannot be presumed. Furthermore, as noted by Moran (1992), not assessing the validity was a violation of APA Standards (AERA, APA, & NCME, 1985).

Conclusion for the WLA. Overall, the WLA has stimuli and a scoring system that do not promote the differentiation of "expert" and "poor" writing qualities, as the use of word ratios was a poor method for determining the quality of written script. Grill and Kirwin used stimuli in the WLA that did not meet any of the stimulus criteria, except that the stimuli were photographic. While the WLA is a direct measure of written expression, the validity of the WLA has been vehemently contested. This measure, thus, met only one criteria proposed by Cole et al. (1997).

The WET

The Written Expression Test (WET; Johnson & Hubly, 1979) is another battery used to assess written expression directly for school aged children in the first to sixth grades. The test was developed to enable easier use for teachers, as it only takes about five minutes to administer and may be given in either an individual or group format. However, the WET was not based upon any theoretical model of written language (Gregg, 1989).

Stimulus criteria. The stimulus is a color photograph, and the desired intent of the photograph was cited as a picture "picked because it elicited good compositions at all grade levels" (Johnson & Hubby, 1979). However, the specific qualities of what a "good" composition contains were not detailed by the authors. The picture depicts a scene where three young school children are engaged in an academic activity. Thus the prompt did meet two of the four stimulus criteria - it is a photograph and contains at least two characters. Although the stimuli may be viewed as interesting, they are hardly novel and could not justifiably be classified as novel and interesting. Also, the stimulus does not depict a conflict among the school children. Overall, the stimulus here does well, but could be more thematically developed in order to represent a novel scene depicting a state of conflict.

Discriminating ability of the scoring criteria. The scoring of the WET contains four sections: Productivity, Mechanics, Handwriting, and Maturity. The criteria are defined mainly by word quantities. None of these criteria allow differentiation of "expert" and "poor" writing qualities per the same problem noted with the WLA. The WET does not meet this criteria.

Modality of assessment and psychometric properties. The WET allows multi-paragraph responses from writers and was therefore classified as a direct measure of written expression. Hence, the WET met the modality of assessment criterion.

The psychometric properties of the WET are not appropriately presented in the manual (Haber, 1989). The only reliabilities shown in the manual are interrater reliabilities. While the numbers are high, the method of assessment is distressing. Only the authors were used in the assessment of interrater reliability, and thus it is not generalizable to the typical professional using the measure. While other reliability studies on the measure were discussed in the manual, they were reported as incomplete. The only validity data presented in the manual was a convergent analysis with the *Comprehensive Test of Basic Skills* (*CTBS*) reading comprehension and vocabulary. A more appropriate analysis would have included another measure of written expression (Gregg, 1989).

Conclusion for the WET. The WET is a direct measure of written expression that meets few of the criteria suggested by Cole et al. (1997). The WET does contain a color photographic stimulus which meets two of four stimulus criteria and is a direct assessment of written expression. However, the scoring criteria do not allow differentiation of "expert" and "poor" writing qualities. Finally, the psychometric properties of the WET are partially unknown due to a failure to include such critical data in the manual and questionable procedures for the analyses reported.

The TEWL and TEWL-2

The Test of Early Written Language (TEWL; Hresko, 1988) is an indirect measure of early writing abilities for children from age 3 to 7 (Wheeler, 1992). The Test of Early Written Language Second Edition (TEWL-2; Hresko, Herron, & Peak, 1996), released during the creation of this article, assesses children from ages 3-0 to 10-11 years. The TEWL and TEWL-*2* are companions to the TOWL-2 and TOWL-3, respectively.

Stimulus criteria. The TEWL contains some auditory stimuli questions and some sections with visual prompts (black and white line drawings). Stimuli are presented to children to assess skills relating to transcription, conventions of print, communication, creative expression, and recordkeeping. Some of the items contain visual stimuli, however the drawings are generally bereft of any thematic content - for example, a drawing of a raincoat or a pencil. Therefore, the TEWL stimuli failed to consistently have the specified type of stimulus (color photograph), contain at least two characters, depict a novel and interesting scene, or depict a state of conflict. Thus none of the stimulus criteria were met by the TEWL.

The TEWL-2 contains separate indirect and direct assessment sections. For the direct assessment, several different pictorial prompts are available for children ages 5-0 to 6-11, and children ages 7-0 to 10-11. None of the stimuli are photographic - they are all black and white line drawings. Thus, the criteria for the type of stimulus was not met by the TEWL-2. All of the pictures from the TEWL-2 do contain more than two characters. Although the stimuli may have been interesting, they were too far from novel to be classified as novel and interesting. Further, the stimuli did not consistently present a state of conflict.

Discriminating ability of the scoring criteria. TEWL items assess five different writing areas: transcription, conventions of print, communication, creative expression, and recordkeeping. The items contained in communication and creative expression contain items which differentiate "expert" and "poor" writing qualities. However, the scoring system combines all of the different writing areas into a single factor (the Written Language Quotient). Therefore, the TEWL does not contain a normed scale which differentiates the "expert" and "poor" writing qualities.

The TEWL-2 separates scoring for the indirect and direct assessments into a Basic Writing Quotient and a Contextual Writing Composite, respectively. The indirect items measure an understanding of picture vocabulary, sentence structure, and other basal writing skills. These items do not differentiate "expert" and "poor" writing qualities. Contextual writing, the direct assessment section of the TEWL-2, does contain items which assess "expert" and "poor" writing qualities. However, the inclusion of other items in the Contextual Writing Quotient excluded this scale from being normed as a scale capable of differentiating "expert" and "poor" writing qualities. Thus, the TEWL-2 does not contain a normed scale that differentiates "expert" from "poor" writing qualities.

Modality of assessment and psychometric properties. The TEWL is an indirect measure of written expression, individually administered to children from ages 3 to 7. The TEWL-2 contains both a direct and indirect assessment of written expression to be individually administered to children ages 4-0 to 10-11.

Reliabilities of the TEWL were reported for stability and internal consistency. Stability measures ranged from .94 to .97, yet only the oldest children were used for the reliability assessment. The generalizability of the stability to younger subjects is suspect. Internal consistency was measured by what the author called coefficient alpha. However, as the scoring criteria are dichotomous, the formula used was actually Kuder-Richardson (as the coefficient alpha formula becomes the Kuder-Richardson formula when dichotomous items are used). These coefficients were sound, ranging from .74 (for ages 4-0 to 4-5) to .95 (for ages 3-0 to 3-5).

Content validity was assessed by analysis of item difficulty and discriminability. While concurrent and criterion-related validity were assessed, the generalizability was low since the author included only 7 year olds in these analyses. Although the reliability estimates appear to be sound and within the normal range for indirect written measurements, no interrater reliability statistics were presented by the author. As previously mentioned, the assessment of interrater reliability for written expression assessment is crucial. Many written expression assessments have questionable interrater reliability and the omission of such an analysis left serious unanswered questions. Also, the validity estimates were of limited interpretative use due to the small range of ages used in these analyses.

Internal consistency for the TEWL-2 was assessed for the Basic Writing Quotient, Contextual Writing Quotient, and the Global Writing Quotient. Cronbach's (1951) coefficient alpha was used to assess the level of internal consistency for the subtests. Alphas for the Basic Writing Subtest ranged from .94 (3 year olds) to .99 (9 year olds). However, the use of coefficient alpha for the Contextual Writing and Global Writing Quotients violated the assumption of independence for coefficient alpha. As noted in the prior psychometric discussions, internal consistency measures are inappropriate when a single stimulus is used. Thus, the Contextual Writing Quotient should not have been assessed for internal consistency. Whereas the Contextual Quotient is a part of the Global Writing Quotient, internal consistency should not have been assessed for the Global Writing Quotient either.

Test-retest reliabilities were also calculated for the three Quotients. The test-retest studies were administered in four separate clusters, varying either age, residency, or both for the subjects in these studies. Administrations were given with 14 to 21 days separation between the initial and secondary testing. All administrations also varied the administration of forms A and B (parallel forms of the TEWL-2) during these administrations – some students received A-A, some B-B, and some A-B. Basic Writing Quotient correlations ranged from .83 to .91. Contextual Writing Quotient correlations ranged from .82 to .88, and Overall Writing Quotient ranged from .91 to .94.

Although the coefficients attained in the test-retest study met all the sound psychometrics criterion, methodological flaws existed during the collection of these data. First, only three cities, very similar in geographical location, were used in the test-retest study: Dallas, Texas; Kansas City, Kansas; and Baton Rouge, Louisiana. Generalization of this study to other regions may be limited. Second, two age groups from the normative sample were not included in these studies; ages 5-0 to 6-11, and 10-6 to 10-11 were missing from the four different clusters assessed without any explanation in the manual. Whereas the coefficients for other age groups were relatively cohesive, and thus it may be assumed that these missing age groups will not deviate radically from the attained correlations, this hypothesis was not confirmed by the authors of the TEWL-2. Third, the authors combined the test-retest analyses with the form equivalency (forms A and B) comparison. Yet, no correlation between the forms with the total normative sample was reported in the manual. While the correlation of the two-forms inside of a test-retest format yielded a desirable correlation, the authors should have provided more comprehensive data on the correlation between these forms. Fourth, too few subjects were assessed in the test-retest analyses for the purposes of their study. For instance, 30 subjects were used in Baton Rouge. These subjects were broken into three groups to assess the test-retest correlations across the possible form combinations (A-A, B-B, A-B). At best, this leaves only 10 subjects per group. Normality, a necessary assumption (Vogt, 1993, p. 155) when assessing correlations (Hamilton, 1992, p. 42), was problematic for the Baton Rouge sample. Although Keppel (1991, p. 97) noted that normality is robust to violations, this only holds when sample sizes are greater than 12 (Bradley, 1980a; Bradley, 1980b; Clinch & Keselman, 1982; Tan, 1982). Ultimately, the violation of normality was compensated by greater sample sizes in the other locations, as these provided replication. Normality estimates for these other locations should have been provided given the violation in the Baton Rouge sample.

Interrater reliability for the Contextual Writing Quotient of the TEWL-2 was very high, especially for a direct measure of written expression. While some methodological problems occurred in the assessment of TEWL-2's interrater reliability, the estimates are still quite admirable. Six raters assessed 25 protocols in the interrater reliability analysis. Table 6.4 of the TEWL-2 manual (Hresko et al., 1996, p. 60) presents the results of the aforementioned analysis. Pearson *r*s for the pairs of raters ranged from .92 to .99, with an average *r* of .95.[4] Although the authors present a complex paragraph explaining why means should be assessed along with correlations in an interrater reliability assessment, they did not use an intraclass correlation which would have provided such an analysis (see Shrout & Fleiss, 1979). The most troubling aspect of the interrater reliability analysis for the TEWL-2 regarded the discernment of the protocols used for the analysis. Whereas the manual first states that "The tests were drawn at random from the overall sample of children . . ." (Hresko et al., 1996, p. 60), they continue by stating that ". . . [the tests] were examined to assure that a range of ability was measured." No further clarity was given regarding this process. Demographic variables for the subjects used in the interrater reliability analysis should have been provided in the manual. Increased heterogeneity, as noted previously, spuriously inflates correlations. Despite these methodological flaws, the true reliability coefficients are likely still to be impressive for a direct measure of written expression.

Construct Validity for the TEWL-2 was assessed by examining test item bias, specifically Differential Item

Functioning (DIF). The authors of TEWL-2 provide a concise, yet adequate, explanation about what DIF is, and how it can be assessed. The authors presented arguments for not using delta scores in DIF (see Camilli & Shepard, 1994), and then proceeded to use this method. However, the authors should be commended for their attempt to offer some analysis of DIF in the complex area of written expression. Difficulties faced the authors. According to Osterlind (1983, p. 12), the concept of unidimensionality underlies all item bias analyses. Yet, as noted by Shepherd and Uhry (1993), reading and writing are heavily intertwined. Using a non-parametric DIF analysis may have been beneficial for the TEWL-2 given the violation of unidimensionality in the test (for example, see the SUBTEST; Stout & Roussos, 1995).

The TEWL-2 was found to have high reliability and validity, as presented in the manual (Hresko et al., 1996). However, several of the analyses were inappropriate or conducted haphazardly. Confirmation of the results reported in the TEWL-2 manual should be conducted through replication. According to the criteria by Cole et al. (1997), the psychometric studies were not conducted with sound methodology and, therefore, did not meet the criterion for psychometric soundness.

Conclusion for the TEWL and TEWL-2. The TEWL did not meet any of the criteria for valid stimuli, did not have a normative scale to differentiate writing ability, did not contain a direct measure of written expression, and its authors did not properly assess the TEWL's psychometric properties. Thus, the TEWL did not meet any of the criteria proposed by Cole et al. (1997).

However, the TEWL-2 was an impressive revision of the TEWL. Some visual stimuli are available for the Contextual Writing Subtest. Although these stimuli only meet the criterion for containing at least two characters, they were also interesting depictions (just not novel). The TEWL-2 did not contain a normed scale which differentiates writing ability, though some of the items in the Contextual Writing Subtest were able to differentiate. Also, the TEWL-2 contains both direct and indirect modalities for written expression assessment. Psychometric properties of the TEWL-2 reported in the manual were quite impressive. Although they contain many methodological flaws, the flaws were not severe. TEWL-2 psychometric properties did not meet Cole et al.'s psychometric criterion due to the difficulties with methodology. Yet, replication of the TEWL-2's psychometric studies will likely provide desirable results. Therefore, although the TEWL-2 only met the criterion for containing a direct modality of assessment, it is a marked improvement over the original TEWL.

The IWI

The Informal Writing Inventory (IWI; Giordano, 1986) is a direct measure of written expression for children from the third to twelfth grades. The IWI is intended to diagnose, categorize error types, and provide remediation for common errors through exercises provided with the test. Although most recent measures of written expression focus on the quantitative extent of ability, Ruben (1992) considered the IWI a regression to remediation models because the IWI focuses on the occurrence of errors.

Stimulus criteria. Fourteen photographic stimuli are available for the elicitation of a written response. Whereas many different depictions are portrayed in the pictures, a general theme of "mild anxiety" is portrayed in all 14 photos (Ruben, 1992). The photos contained in the test do meet the stimulus criterion of portraying a conflict state. Stimuli do not consistently portray two or more characters, or novel and interesting depictions. Unfortunately, the consistency of conflict is not held throughout all possible stimuli as the test does allow examinees to draw their own picture. Therefore, while the photos provided with the test met two of the four stimulus criteria, the test methodology, which allows an examinee to create their own stimulus, may preclude all stimulus criteria from being met (depending on the drawing created by the examinee).

Discriminating ability of the scoring criteria. Specifically, formation (handwriting), grammar, communication and total errors are tabulated after the subject writes about a pictorial stimulus. Communication errors, as suggested through its nomenclature, are purported to assess "expert" and "poor" writing qualities. Yet, the ambiguous nature of the manual (see Della-Piana, 1992) suggested that this scale is only to be comprised of errors of formation and grammar, and errors of logic. Hence, none of the scales in the IWI discriminate between "expert" and "poor" writing abilities.

Modality of assessment and psychometric properties. The IWI does not restrict the length of a written response by an examinee and thus has been classified as a direct measure of written expression. Therefore the modality criterion was met by the authors of the IWI.

Since the IWI was introduced as an informal assessment of writing ability, it contains no reliability or validity information. However, both reviews of the IWI in the 1992 Mental Measurements Yearbook discuss copious interpretation problems with the scoring that would likely lead to poor reliability (Della-Pina, 1992; Ruben, 1992).

Conclusion for the IWI. Overall, while many pictorial stimuli are included with the IWI, none of the stimulus criteria are consistently met across all possible administrations. The lack of consistency was mainly a product of an allowance for examinees to create their own stimulus. This test is a direct measure of written expression with no ability to differentiate "expert" or "poor" writing qualities, and it concomitantly has poor scoring criteria likely leading to low reliability and validity, though no data are presented by the authors of the IWI. Overall, the IWI only met one criterion (it is a direct measure of assessment) proposed by Cole et al. (1997).

The IAS

The Integrated Assessment System (IAS; Farr & Farr, 1990) contains a direct assessment of language arts skills for children grades one through eight. Farr and Farr developed the IAS to assess abilities and skills in the language arts, particularly suited for schools using whole language and literature based instruction approaches (Farr & Farr, 1991).

Stimulus criteria. The IAS presents various written prompts, containing either unique stories or common childhood literature, that requires a written response from the child. Whereas the prompts are not photographic, or even visual, they did meet the criterion of a novel and interesting depiction. The story prompts in the IAS are consistently interesting and portray events with at least subject novelty. The stories do not, however, consistently portray two or more characters, or a state of conflict. Overall, stimuli from the IAS only met one of the four stimulus criteria.

Discriminating ability of the scoring criteria. Three dimensions are graded for each response, ranging from 1 to 4 points. Response to Reading assesses the comprehension of the passage read, including accuracy and quantity of information recapitulated by the child. Management of Content assesses the child's organizational ability, including maintaining focus and leading to resolution. This facet should moderately discriminate between qualities of "expert" and "poor" writers. Command of Language is the last dimension assessed in the IAS and it relates to grammatical and lexicological usage.

Modality of assessment and psychometric properties. The IAS was considered to be a direct measure of written expression, as examinees are allowed to provide responses longer than a single paragraph. The modality of written expression assessment, therefore, was met by the authors of the IAS.

Pearson and Spearman reliability correlations are presented in the manual for interrater reliability (Farr & Farr, 1991). Most of the correlations are in the .80s and .90s - strong correlations for a direct measure of written expression. Yet, the correlations suffer from generalizability problems. The raters in the interrater reliability assessment were "trained readers at The Psychological Corporation's Writing Assessment Center . . . [who have had] . . . previous experience with large-scale writing assessment and had worked on numerous related projects" (Farr & Farr, 1991, p. 27). The typical user of an achievement test is not likely to have as much experience in writing assessment.

Data reported in the manual indicate that convergent and discriminant validity correlations range from .81 to .94. Again, problems existed in their methodology. The authors used Campbell and Fiske's (1959) multitrait-multimethod procedure. Though the authors' explanation of this procedure in the manual is well-grounded, deviations from Campbell and Fiske's recommendations occurred in Farr and Farr's (1991) assessments. As summarized by Campbell and Fiske (p. 103), methods and traits need to be maximally dissimilar. Yet, Farr and Farr used Response to Reading, Management of Content, and Command of Language as their traits, and independent scorers were considered different methods. Fiske and Campbell (1992) have expressed their frustration regarding the improper use of the dissimilar methods and traits. The IAS used traits which all come from a language based domain and a method that, though containing the word independent in its description, is anything but dissimilar. The use of many judges is a poor interpretation of dissimilarity amongst methods, as rating by an expert is a single method regardless of the number of raters used.

Although the authors of the IAS claim that the test has a combined ability to assess both reading and writing skills together, in fact, the IAS does not have a way of differentiating between reading and writing abilities. Factor analysis and convergent validity should have explored this area.

Conclusion for the IAS. Summarily, this measure met some of the criteria set forth by Cole et al. (1997): scoring is likely to allow some differentiation of "expert" and "poor" writing, the assessment is conducted in a direct modality, and initial psychometric properties appear to be reasonable. However, the prompts only met the criterion of a novel and interesting depiction. Additionally, further analyses of the psychometric properties are needed.

Summary

The summary of comparisons is presented in Table 1. The results of the comparison of written expression measures were that none of these measures met all of the criteria suggested by Cole et al. (1997). None of the measures met all the stimulus criteria, although five measures met three of the four stimulus criteria. The TOWL-3 allows discriminability for "expert" and "poor" writing qualities. The OWLS and IAS also contain aspects in their scoring system which allow for some discriminability among writing qualities. However, the other nine measures did not allow for such differentiation. Furthermore, most of the measures had questionable psychometric analyses. In fact, only the WJ-R/A and OWLS met this criterion.

In order to provide some means for comparisons among the tests, three categories have been created. Categories were created based on the number of criteria a test met. The stimulus criteria were counted individually, therefore a total of seven criteria were able to be rated. Good tests were those that met at least four of the seven criteria. Moderate tests were those containing two or three of the criteria. Finally, poor tests were those which met fewer than two criteria. The TOWL-2 and TOWL-3 were the only tests rated as good. The TOWL-3 was specifically remarkable for its ability to differentiate "expert" and "poor" writing qualities, and its ability to attain two of three stimulus criteria, almost attaining a third. The PIAT-R, OWLS, WLA, TEWL-2, WET, IAS, and IWI were rated as moderate tests. All of these tests, except for the OWLS, suffered from poor assessment of their respective psychometric properties. Last, the WIAT, WJ-R/A, and TEWL were rated as poor tests. These measures should be carefully considered by test examiners for their appropriateness given the tests' lack of theoretical development.

In conclusion, new work is needed that incorporates theoretically based research, the criteria set forth by Cole et al. (1997), and the integration of other current theoretical research on written expression, to produce a psychometrically sound, highly useful test of written expression. However, it should be noted that Cole et al. postulates have not been empirically validated as a unit. Further research should attempt to assess the veracity of this system. The postulates proposed by Cole et al. are not the final word in written expression theory; instead, they are the important first rung of the theoretical ladder.

Table 1
Summary of Cole et al.'s (1997) Criteria Compared to Existing Measures of Written Expression

	Criteria						
	Stimulus Criteria						
Test	Type of Stimulus	Two Characters or more	Novel and interesting depiction	State of conflict	Scoring discriminability	Direct or indirect response	Sound psychometrics[a]
PIAT-R	LD	Yes	No	No	No	Direct	No
WIAT	A & T	No	No	No	No	Direct	No
WJ-R/A	A	No	No	No	No	Indirect	Yes
TOWL-2	LD & A	Yes	Yes	Yes	No	Both	No
TOWL-3	LD & A	Yes	Yes	Yes	Yes	Both	No
OWLS	A & T	No	No	No	Partially	Indirect	Yes
WLA	CP & T	No	No	No	No	Direct	No
WET	CP	Yes	No	No	No	Direct	No
TEWL	A & LD	No	No	No	No	Indirect	No
TEWL-2	LD	Yes	No	No	No	Both	No
IWI	CP	No	No	No[b]	No	Direct	No
IAS	T	No	Yes	No	Partially	Direct	No

Note. When a measure contained many stimuli the criteria for stimuli must have been consistently met in order to have passed the criteria.

LD = line-drawing; CP = color photo; A = Auditory; T = Text.

[a] For a test to be labeled as psychometrically sound it must exceed a reported reliability of .80 and have had sound methodology in the psychometric assessments.

[b] Whereas the stimuli from the IWI do create a state of conflict in all of the stimuli provided with the test, the allowance for examinees to create their own stimulus precludes this test from consistently maintaining conflict in all possible prompts.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing.* (6th ed.). New York: Macmillan.

Anderson, P. L. (1989). Productivity, syntax, and ideation in the written expression of remedial and achieving readers. *Journal of Reading, Writing and Learning Disabilities International, 4,* 115-124.

Auchter, J. C., & Hatch, M. (1990, March). *Evaluating multiple writing literacies through multiple choice testing and direct assessment.* Paper presented at the 1990 Annual Conference of the National Testing Network on Writing, New York, NY.

Benes, K. M. (1992). Peabody Individual Achievement test -Revised. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 649-652). Lincoln, NE: Buros Institute of Mental Measurements.

Benton, S. L. (1992). Test of Written Language - 2. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 979-981). Lincoln, NE: Buros Institute of Mental Measurements.

Bereiter, C. (1980). *Development in writing.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Berninger, V. W., Mizokawa, D. T., Bragg, R., & Cartwright, A. C. (1994). Intraindividual differences in levels of written language. *Reading and Writing Quarterly: Overcoming Learning Difficulties, 10,* 259-275.

Bradley, J. V. (1980a). Nonrobustness in classical tests on means and variances: a large-scale sampling study. *Bulletin of the Psychonomic Society, 15,* 275-278.

Bradley, J. V. (1980b). Nonrobustness in z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society, 16,* 333-336.

Breland, H. M., & Gaynor, J. L. (1979). A comparison of direct and indirect assessment of writing skills. *Journal of Educational Measurement, 16,* 119-128.

Burtis, P., Bereiter, C., Scardamalia, M., & Tetroe, J. (1983). *The development of planning in writing.* Chichester, England: John Wiley.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* (Vol. 4). Thousand Oaks, CA: Sage.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carrow-Woolfolk, E. (1996). *Oral and Written Language Scales.* Circle Pines, MN: American Guidance Service.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics, 7,* 207-214.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 115-159.

Cole, J. C., Muenz, T. A., Ouchi, B. Y., Kaufman, N. L., & Kaufman, A. S. (1997). The impact of the pictorial stimulus on the written expression output. *Psychology in the Schools, 34*(l), 1-9.

Cooper, P. L. (1994). The assessment of writing ability: A review of research (Vol. ED 250332,). Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Della-Piana, G. (1992). Informal Writing Inventory. in J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 410-411). Lincoln, NE: Buros Institute of Mental Measurements.

Dunn, L. M., & Dunn, L. (1981). *Peabody Picture Vocabulary Test - Revised.* Circle Pines, MN: American Guidance Service.

Englert, C. S. (1990). *Unraveling the mysteries of writing through strategy instruction.* New York: Springer-Verlag.

Englert, C. S., Raphael, T. E., Anderson, L. M., Gregg, S. L., & Anthony, H. M. (1989). Exposition: Reading, writing, and the metacognitive knowledge of learning disabled students. *Learning Disabilities Research, 5,* 5-24.

Englert, C. S., & Thomas, C. C. (1987). Sensitivity to test structure in reading and writing: A comparison of learning disabled and non-learning disabled students. *Learning Disabilities Quarterly, 10,* 93-105.

Farr, R., & Farr, N. (1990). *Integrated Assessment System: Examination Kit:* Psychological Corporation.

Farr, R., & Farr, N. (1991). *Integrated Assessment System: Preliminary Technical Report:* Psychological Corporation.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve the problem. *Psychological Bulletin, 112*(2), 393-395.

Fitzgerald, J. (1987). Research on revision in writing. *Review on Educational Research, 57*, 481-506.

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*, 365-387.

Giordano, G. (1986). *Informal Writing Inventory*. Salt Lake City, UT: Scholastic Testing Services.

Graham, S., & Harris, K. R. (1987). Improving composition skills of inefficient learners with self-instructional strategy training. *Topics in Language Disorders, 7*, 66-77.

Graham, S., & Harris, K. R. (1989). A components analysis of cognitive strategy training: Effects of learning disabled students' compositions and self-efficacy. *Journal of Education Psychology, 81*, 353-361.

Graham, S., Harris, K. R., & MacArthur, C. A. (1990). Learning disabled and normally achieving students' knowledge of the writing process. Unpublished raw data.

Graham, S., Harris, K. R., MacArthur, C. A., & Schwartz, S. (1991). Writing and writing instruction for students with learning disabilities: Review of a research program. *Learning Disabilities Ouarterly, 14*, 89-114.

Gregg, N. (1989). The Written Expression Test. In J. J. Kramer & J. C. Conoley (Eds.), *The tenth mental measurements yearbook* (pp. 921-922). Lincoln, NE: Buros Institute of Mental Measurements.

Grill, J. J., & Kirwin, M. M. (1989). *Written Language Assessment*. Novato, CA: Academic Therapy Publications.

Guilford, J. P. (1954). *Psychometric methods*. (2nd ed.). New York: McGraw-Hill.

Haber, L. (1989). The Written Expression Test. In J. J. Kramer & J. C. Conoley (Eds.), *The tenth mental measurements yearbook* (pp. 921-922). Lincoln, NE: Buros Institute of Mental Measurements.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman Group.

Hamilton, L. C. (1992). *Regression with graphics: a second course in applied statistics*. Belmont, CA: Duxbury.

Hammill, D. D., & Hresko, W. P. (1994). *Comprehensive Scales of Student Abilities*. Austin, TX: PRO-ED.

Hammill, D. D., & Larsen, S. C. (1983). *The Test of Written Language*. Austin, TX: PRO-ED.

Hammill, D. D., & Larsen, S. C. (1988). *The Test of Written Language -- Second Edition*. Austin, TX: PRO-ED.

Hammill, D. D., & Larsen, S. C. (1996). *The Test of Written Language -- Third Edition*. Austin, TX: PRO-ED.

Harrison, P. L., Kaufman, A. S., Hickman, J. A., & Kaufman, N. L. (1988). A survey of tests used for adult assessment. *Journal of Psychoeducational Assessment, 6*, 188-198.

Hayes, J. R., & Flower, L. S. (1986). Writing research and the writer. *American Psychologist, 41*, 1106-1113.

Holland, P. W., & Thayer, D. T. (1986). *Differential performance and the Mantel-Haenszel Procedure* (Report Number 86-31). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential performance and Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-146). Hillsdale, NJ: Erlbaum.

Hooper, S. R., Montgomery, J., Swartz, C., Reed, M. S., Sandler, A. D., Levine, M. D., Watson, T. E., & Wasileski, T. (1994). Measurement of written language expression. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 375-417). Baltimore, MD: Paul H. Brookes.

Hresko, W. P. (1988). *Tests of Early Written Language*. Austin, TX: PRO-ED.

Hresko, W.P., Herron, S. R., & Peak, P. K. (1996). *Tests of Early Written Language -- Second Edition*. Austin, TX: PRO-ED.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, C. & Hubly, S. (1979). *The Written Expression Test*. Denver, CO: Rocky Mountain Education Systems.

Kaufman, A. S. (1990). *Assessing Adolescent and Adult Intelligence*. Boston, MA: Allyn and Bacon.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement, Comprehensive Form*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test*. Circle Pines, MN: American Guidance Service.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lord, F. M. (1977a). Practical estimations of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lord, F. M. (1977b). A study of item bias using item characteristic curve theory. In N. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swits & Vitilinger.

MacEachron, A. E. (1982). *Basic statistics in the human services*. Austin, TX: PRO-ED.

Markwardt, F. C., Jr. (1989). *Peabody Individual Achievement Test -- Revised*. Circle Pines, MN: American Guidance Service.

Mather, N. (1993). *Administering, scoring, and evaluating the Writing Samples of the WJ-R* (WJ-R Test 27: Writing Samples). Chicago, IL: Riverside Resource Bulletin.

McCutchen, D., & Perfetti, C. (1982). Coherence and connectedness in the development of discourse production. *Text, 2*, 113-119.

Moran, M. R. (1992). Written Language Assessment. In J. J. Kramer & J. C. Conoley (Eds.), *The tenth mental measurements yearbook* (pp. 1047-1050). Lincoln, NE: Buros Institute of Mental Measurements.

Norris, J. A. (1992). Written Language Assessment. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 979-981). Lincoln, NE: Buros Institute of Mental Measurements.

Osterlind, S. J. (1983). *Test item bias*. Thousand Oaks, CA: Sage.

Ouchi, B. Y., Cole, J. C., Muenz, T. A., Kaufman, A. S., & Kaufman, N. L. (1996). Interrater reliability of the written expression subtest of the Peabody Individual Achievement Test Revised: An adolescent and adult sample. *Psychological Reports, 79*, 1239-1247.

Psychological Corporation. (1991). *Wechsler Intelligence Scale for Children -- Third Edition*. San Antonio, TX: Psychological Corporation.

Psychological Corporation (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Psychological Corporation.

Rogers, B. G. (1992). Peabody Individual Achievement Test -Revised. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 979-981). Lincoln, NE: Buros Institute of Mental Measurements.

Ruben, D. L. (1992). Informal Writing Inventory. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 981-982). Lincoln, NE: Buros Institute of Mental Measurements.

Ryan, J. M. (1992). Test of Written Language -- 2. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 981-982), Lincoln, NE: Buros Institute of Mental Measurements.

Scardamalia, M., & Bereiter, C. (1986). *Research on written composition*. (3rd ed.). New York: Macmillan.

Semel, E., Wiig, H., & Secord, W. (1987). *Clinical Evaluation of Language Fundamentals -- Revised*. San Antonio, TX: Psychological Corporation.

Shepherd, M. J., & Uhry, J. K. (1993). Reading disorder. *Learning Disabilities, 2*(2), 193-208.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428 .

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31*, 378-388.

Stout, W., & Roussos, L. (1995). *SUBTEST user manual*. Urbana, IL: University of Illinois.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departure and normality. *Communications in Statistics -Theory and Methods, 11*(2), 485-511.

Vogt, W. P. (1993). *Dictionary of statistics and methodology*. Newbury Park, CA: Sage.

Wheeler, P. (1992). Test of Early Written Language. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 946-947). Lincoln, NE: Buros Institute of Mental Measurements.

Wong, B., Wong, R., & Blakenship, J. (1989). Cognitive and metacognitive aspects of learning disabled adolescents' composition problems. *Learning Disabilities Quarterly, 15*, 145-152.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests -Revised*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psychoeducational Battery -- Revised*. Chicago, IL: Riverside Publishing.

Woodcock, R. W., & Mather, N. (1989). *WJ-R Tests of Achievement - Standard and Supplemental Batteries: Examiner's Manual*. Chicago, IL: Riverside Publishing.

Footnotes

[1] Perfect agreement on what differentiates direct from indirect assessment in written expression is not found. The main argument among those persons who advocate direct assessment is that a free response should more adequately reflect a writer's ability than an indirect response, hence an increase in construct validity. In fact, Berninger et al. (1994) showed that writing at the word level does not predict writing at the sentence or paragraph

level, nor does writing at the sentence level predict writing at the paragraph level. Story cohesion and fluid transitions have been labeled as writing qualities found in "expert" writers (see the introductory section), and many of these are demonstrated during paragraphical transitions. Therefore, it seems inadequate for a direct measure of written expression to contain anything less than a multi-paragraph format.

[2] Liberal qualification of stimuli exhibiting novel and interesting depictions, as well as their ability to generate a state of conflict, was administered. As these theoretical constructs have partial refutability, they are nevertheless subjective. The authors felt it was better to qualify, rather than eliminate, questionable stimuli in the aforementioned regards.

[3] PRO-ED publishes the TOWL-2 and TOWL-3.

[4] The authors did not provide the method for averaging the *r*s in the interrater reliability analyses. However, the Fisher *Z*-score transformation formula appears to have been used, as it provides the mean *r* reported in the manual.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 35-43

Assessing School Work Culture

William L. Johnson
Ambassador University

Karolyn J. Snyder
University of South Florida

Robert H. Anderson
President, Pedamorphosis, Inc.

Annabel M. Johnson
Ambassador University

School culture surfaced in the 1980s as a context for the study of the development of schools. This article reviews a work culture productivity model and reports the development of a work culture instrument based on the culture productivity model. The second-order component analysis shows areas of generalization across the primary components such as continuous improvement, human resource development and group planning, strategic planning and accountability, and collaboration.

School culture surfaced in the 1980s as a framework for the study of the development of schools (Deal, 1987; Deal & Kennedy, 1982; Greene, 1991; Kushman, 1992; Rossman, Corbett, & Firestone, 1988; Snyder & Anderson, 1986). Culture has been defined as the knowledge that is learned, shared, and used by persons within organizations to interpret experience and generate behavior (Spradley & McCurdy, 1996). It is an understanding of "the way we do things around here" and is characterized by shared beliefs and visions, rituals and ceremonies, and networks of communication (Deal & Kennedy, 1982, p.14). Organizational researchers have sought to understand school culture and link the same with educational productivity. It has been noted that the effect of culture on productivity is so powerful that developing a culture which supports school effectiveness is essential to school success (Deal, 1987). Therefore, numerous reform efforts have focused on bringing about changes in existing school cultures (Hocevar, 1994; Miles & Louis, 1990; Rigsby, 1994).

Studies of organizational culture have used both qualitative and ethnographic approaches, as well as quantitative approaches. Rooted in the concept of systems culture, the construct of school work culture is described as a subset of the same. Specifically, it refers to the collective work patterns of a system (or school) in the areas of systemwide/schoolwide planning, professional development, program development, and assessment of productivity, as perceived by its staff members (Snyder, 1988). This generalization is derived from the literature that schools can have a culture that either supports or hinders educational excellence and productivity and that positive school culture is associated with effective schools (Deal, 1987; Sergiovanni, 1987; Sweeney, 1987).

In a massive nationwide study, Chubb and Moe (1990) randomly sampled 500 schools. Some 10,000 students participated in the testing and surveys, and 12,000 teachers provided in-depth information about decision making, classroom environment, and their perceptions of the problems in their schools. In addition, the principals and administrators in all of the schools were surveyed. The results showed that attending an effectively organized high school is worth at least an extra year's achievement over the course of a high school career. The authors found that a clear sense of purpose,

---

William L. Johnson is chair of the education, psychology, and family studies department, and dean of admissions and financial aid, at Ambassador University in Big Sandy, Texas. Karolyn J. Snyder is a professor in the educational leadership department at the University of South Florida, Tampa, Florida, where she also serves as director of the school management institute. Robert H. Anderson holds an eminent scholar chair at the University of South Florida. He is also President of Pedamorphosis, Inc. in Tampa, Florida. Annabel Johnson is a professor of family studies in the education, psychology, and family studies department at Ambassador University, Big Sandy, Texas. Correspondence regarding this article should be addressed to William L. Johnson, Department of Psychology and Education, Ambassador University, Big Sandy, TX 75755 or by e-mail, william_johnson@ambassador.edu.

leadership, professionalism (treating teachers as professionals), and high expectations for academic work were what really seemed to matter. Overall, the schools seemed to work like a professional team. The researchers found that the most important determinant of what students gain in high school was the students' individual aptitude. But the second most powerful predictor of achievement gains in high school was effective school organization (Brandt, 1990-1991). The whole quality movement has taken form by focusing on the relationship of work culture and its effect on organizational productivity.

Based on this background, the purpose of the present study was to use second-order principal components analyses to assess work culture. A second-order factor analysis will incorporate an additional level of analysis by showing how the first-order factors group into higher-order factors. This is important in assessing the global components of work culture.

Managing Productive Schools

During the past decade, Snyder and Anderson (Snyder, 1988; Snyder & Anderson, 1986) implemented a leadership training program known as Managing Productive Schools (MPS) in Florida, Minnesota, and Virginia. The Minnesota State Legislature has adopted the MPS job dimensions (work culture) as the licensing rules for principals. In Israel, Professor Tamar Horowitz (from Ben Gurion University in the Negev) has been invited by the Israel Department of Education to design a schoolwide assessment system based on the MPS job dimensions. In the late 1980s, representatives of the Government of India requested that they be allowed to distribute these findings to the leading educators in India. The program is based on the research base noted above and also on a systems approach to organizational development. That is, all dimensions of the organization are viewed as interdependent features to enable the system to achieve its purposes and goals. Following is a brief review of the school work culture model.

Dimension 1: Schoolwide Planning. As Rigsby (1994) noted, "This constancy of purpose, restated and reinforced by top level management fosters a culture of cooperation, teamwork, innovation, and a commitment to continual improvement and customer satisfaction" (p. 5). Perkins (1994) wrote that the work of teams is the wave of the future as collaboration and a sense of community between and within all departments and levels of the organization have replaced the working mode of isolation. Grade level teams, curriculum committees, the school site council, *ad hoc* problem solving committees, and staff meetings contribute significantly to a professional culture in the school (Chrispeels, 1992). Peters and Austin (1985) found that the intensity of management's commitment to organizational goals is the chief difference between great and not-so-great organizations.

Dimension 2: Professional Development. Professional development plans that are linked to organizational goals have the power to enhance individual and group performance, and that of the school as well (Carneval, 1989). With little time for development, Chrispeels (1992) stated that teachers do not have the opportunity to develop their teaching skills. They may also lack self-confidence and feel their self-esteem threatened. However, teachers felt that barriers were being broken down through schoolwide staff development programs and that staff development was critical to school improvement. Structurally, work groups become learning centers for teachers as they share, plan, and critique programs or tasks together (Larson & LaFasto, 1989).

Dimension 3: Program Development. The purpose of program development is to solve specific problems and solve learning challenges. However, the top leadership of educational institutions are those who must provide the organization with the values and guiding philosophy inherent in a quality culture (Hocevar, 1994). Interestingly, commitment to change seemed more prevalent among staff members in more effective schools than in less effective schools (Chrispeels, 1992). Teacher collaboration is evident in a successful school (Kushman, 1992).

Dimension 4: School Assessment. Accountability systems drive assessment activities in productive organizations. Goal-based assessments are the most effective in altering individual and organizational performance. All systems need feedback to remain viable, and feedback requires information about accomplishments in relation to the purposes, goals, and output of the system or organization (Chrispeels, 1992). Those closest to the work have the greatest opportunity to understand the work and know what needs to be done for improvement (Stratton, 1991).

The expansion of the literature base about organizational and human productivity indicates that administrators and teachers together must assume responsibility for students' achievement patterns to change. Smylie and Denny (1990) noted that change must be grounded in local discretion and in decision making that involves teachers as participants. The existence of formal team structures is related directly to increases in the degree of teacher involvement in decision making (Blase, 1993). The role of the principal, for example, has changed from keeping teachers in their rooms to leading teachers in areas such as budget, personnel, curricular, and instructional considerations (Walker & Peel, 1993).

The literature on productive educational, social, and business organizations continues to affirm that employee involvement is essential to the very survival of an organization. Resources, information, opportunity, and support are vital materials that fuel organizational productivity (Johnson & Snyder, 1989-1990). A typical production model might divide the school year into three parts: planning (September and October), development (November through April), and evaluation (May and June). Planning activities might include schoolwide goal setting and work group and individual staff performance planning. Developmental activities might include staff development, clinical supervision, work group development, and quality control activities. Program development might include instructional program and resources development. Productivity assessment would include assessing achievement for students, teachers, work groups, and the school itself. The assessment findings would then serve to direct the feedback and feedforward planning and development activities for the next academic year.

The work culture model was based on an in-depth study of the literature on productive organizations and work cultures in business and education; over 400 studies were reviewed (Snyder, 1988; Snyder & Anderson, 1986). Included within the four subscales are 10 smaller logical clusters (dimensions): goal setting, work group performance, individual staff performance, staff development, clinical supervision, work group development, instructional program development, resources development, quality control, and assessment. The implementation of these work-culture dimensions defines a school production model. See Appendix A for an outline of this model.

Method

Participants

The total sample of subjects (n=925) were from 112 Florida schools representing 41 of the 67 school districts in Florida. The ratio of teachers to principals was approximately four to one. Participants were asked neither their ages nor their gender; however, they were typical elementary educators representing more than half of the school districts in Florida. Each subject in the sample was sent a survey instrument with directions and a machine-scorable answer sheet. The data were collected by mail.

Materials

School work culture was operationalized on the *School Work Culture Profile* (SWCP) with 60 statements pertaining to existing work practices in a school. A five-point Likert scale ranging from *strongly disagree* to *strongly agree*, with a midpoint of *undecided*, was used to rate each item. The 60 items represented four subscales of 15 items each. The domains were titled schoolwide planning, professional development, program development, and school assessment. Included within the four domains were the 10 smaller clusters noted earlier.

Instrument

In the 1980s, the authors were conducting administrative workshops in the United States and Canada. During a seminar in Prince George, British Columbia, superintendents asked if much of the workshop information might be incorporated into an assessment instrument that school administrators could use in their schools. After examining the research base for the productivity model being discussed, 100 research-based subset skills were translated into a diagnostic instrument. The instrument was piloted in workshops over the next year, and in 1984 a revised instrument was field tested in Missouri, Maryland, and Florida.

In 1987, the instrument was edited and reorganized. Items were assigned randomly to the instrument, and directions were written to allow for use with a machine-scorable answer sheet. Finally, two mailings were sent to a nationwide panel of 17 experts in the field who were asked to evaluate the instrument for language clarity and item relevance. These content validity surveys led to the current *School Work Culture Profile* instrument.

The instrument measures were submitted to reliability testing in the summer of 1987. A sample of 46 elementary school teachers in Pasco County, Florida responded to the items. The Cronbach alphas were strong indicators of reliability. Several items were dropped or modified, and one subset of statements was moved from the staff development subscale to the assessment subscale. These refinements resulted in alpha reliability estimates of .82 to .95 on the schoolwide planning, professional development, program development, and school assessment subscales and a composite scale alpha of .95.

The SWCP was tested using two different reliability samples. Two classes of graduate students in education (n=46) took the SWCP in the fall of 1987. Alphas for the four subscale measures were between .88 and .93, and the alpha for the composite was .97. A second sample of 50 elementary school teachers in Lee County, Florida participated in a test-retest study with a two-week delay time in the spring of 1988. A test-retest Pearson correlation coefficient of .78 was attained.

Results

We used the SAS principal components program (SAS Institute, Inc., 1986) to examine the factorial (construct) validity of the instrument. A relevant question pertains to the researchers' use of principal components versus principal factor analysis. Will different factors emerge if 1.00s are put into the main diagonal rather than communalities? In an analysis, the number of variables affects the degree of difference between the two methods. For example, with 10 variables, 10% (10/100) of the entries involve the diagonal of the correlation matrix, but, with 60 variables, 1.6% (60/3,600) of the entries are in the diagonal. Gorsuch (1983) stated when there was a large number of variables having moderate loadings, the difference between the two analyses was negligible. Nunnally (1978) wrote, "It is very safe to say that if there are as many as 20 variables in the analysis, as there are in nearly all exploratory factor analyses, then it does not matter what one puts in the diagonal spaces" (p. 418). Velicer and Jackson (1990) noted that the choice of method was unlikely to result in empirical or substantive differences. This reasoning constituted the justification for performing a principal components analysis rather than a principal factor analysis.

Determining the number of factors to extract from the correlation matrix is a fundamental decision in any analysis (Thompson & Borello, 1986). Many researchers follow the recommendations of Guttman (1954) and extract all factors with eigenvalues greater than one. We used the eigenvalue criterion for this study since the number of respondents was greater than 250 and the mean communality was approximately 0.60 (Stevens, 1986).

Initially, we performed a first-order principal components analysis (Pedhazur & Schmelkin, 1991; Stevens, 1986). Individual questions were retained if they had a pattern/structure coefficient greater than or equal to $|0.40|$. The first-order principal components analysis yielded ten factors. The prerotation eigenvalues for the components ranged from 1.02 to 20.38.

One result of the first-order principal components analysis was a matrix of correlations among the factors. The interfactor correlation matrix can be factored just as the 60 x 60 variable matrix can be. The decision to extract second-order factors was driven by the finding that the first-order factor correlation matrix had numerous noteworthy correlations, suggesting a first-order oblique solution as well as a second-order result. Very often in research, the value is set at 0.4 in absolute magnitude. Items were included if they had pattern/structure coefficients greater than or equal to 0.40 in absolute value. Gorsuch (1983) noted the similarity of procedures for both higher-order and primary analyses; therefore, the authors used the eigenvalue criterion in determining the number of higher-order factors. Gorsuch (1983) noted the eigenvalue criterion was an appropriate approach for higher-order analysis.

The 60 x 10 promax rotated first-order factors were postmultiplied by the 10 x 4 varimax rotated second-order factors, and the 60 x 4 product matrix was then rotated to the varimax criterion. This 60 x 4 product matrix was the desired second-order solution. The decision to conduct an orthogonal rotation at any order terminates the higher-order sequence (Loehlin, 1992). The second-order factor matrix was rotated to the varimax criterion because the orthogonal rotation finalized the higher-order sequence. See Table 1 for the second-order factor loadings.

We used the generalized Kuder-Richardson reliability formula, coefficient alpha (Cronbach, 1970), to estimate the reliability of the instrument measures. This formula was appropriate since a scale in Likert format was employed. The Cronbach alphas for the factors (subscales) follow: subscale one .92, subscale two .88, subscale three .44, subscale four .67, and the composite for all questions .94. The alpha values for subscales three and four were not considered major impediments because the reliabilities are a function of the subscale lengths.

Discussion

The second-order factor analysis generated a set of relationships among the 60 items on the *School Work Culture Profile* which reflect several major thrusts for organizational transformation within the quality management literature. We have given the following names to the four second-order factors: Continuous Improvement, Human Resource Development and Group Planning, Strategic Planning and Accountability, and Collaboration. A greater interdependence among logical work culture dimensions has emerged, and this reinforces the systems thinking imbedded within the SWCP. As previously noted, high school student achievement is linked with effective school organization. This study develops that linkage and offers an instrument to test the assertions.

Table 1
Rotated Pattern/Structure Coefficients for Salient Items

Item	Question	Factors 1	2	3	4
1	The school administration and the staff identify goals to improve the school each year.	**0.720**	0.330	-0.042	0.080
2	The staff development program builds the school's capacity to solve problems.	**0.513**	0.242	-0.031	0.375
17	Staff members have opportunities to develop skills for working successfully in a group/team.	**0.400**	0.115	0.134	0.173
18	School evaluation is based on school goals.	**0.409**	0.265	-0.114	-0.167
19	Tasks are identified for accomplishing school development goals.	**0.421**	0.308	-0.035	-0.084
34	Work groups report periodically on progress to the school leadership team.	**0.558**	-0.035	-0.138	-0.007
35	School-wide task forces and committees work to achieve school development goals.	**0.663**	0.231	-0.051	0.019
41	Work group plans are reviewed by the leadership team.	**0.400**	-0.186	-0.300	0.089
49	Work group leaders have opportunities to develop specific leadership skills.	**0.485**	0.255	-0.006	0.103
50	All staff members develop individual performance goals to contribute to school development goals.	**0.637**	-0.344	-0.040	0.085
54	Staff member's share their ideas and concerns for improving work productivity in their work group	**0.544**	0.001	0.120	0.304
55	The school's leadership team helps work groups to succeed.	**0.544**	0.148	-0.007	0.124
57	Individual performance goals for staff members are linked to the school's development goals.	**0.678**	-0.221	-0.044	-0.052
58	Staff members problem solve, plan, and make decisions together in productive ways.	**0.425**	0.260	0.198	0.348
3	Instructional programs are guided by learning objectives.	0.068	**0.442**	-0.038	0.274
6	Staff development programs provide opportunities to learn new knowledge.	0.153	**0.599**	0.062	0.382
10	Parents participate in identifying school goals.	-0.135	**0.471**	-0.055	-0.110
16	Instructional programs facilitate student mastery of learning objectives.	-0.057	**0.440**	0.149	0.150
21	School evaluation includes assessment of student achievement data.	0.140	**0.684**	-0.134	-0.049
26	Students are provided with reinforcement, correctives, and feedback on their performance.	0.108	**0.400**	0.239	0.032
36	Supervision helps teachers to solve instructional problems.	0.233	**0.400**	0.158	0.215
37	Resources are used to meet school goals.	0.278	**0.601**	0.308	-0.017
38	Commonly held beliefs, values and norms are consistent with school development goals.	0.339	**0.434**	0.376	-0.159
42	Parents serve as a resource to the school's instructional program.	-0.179	**0.575**	0.212	-0.098
43	Supervision builds and maintains professional self-esteem.	0.089	**0.435**	0.306	0.064
45	High performance expectations exist for each role group (for example: teachers, counselors).	0.230	**0.481**	-0.019	0.138
46	Supervision reinforces strengths in current job performance.	0.080	**0.451**	0.252	0.151
47	Community resources are used in the school's instructional programs.	0.033	**0.558**	0.154	0.004
52	The school's budget reflects prioritized school goals.	0.204	**0.453**	0.124	-0.266
4	Work groups (committees, department teams, grade level groups, etc.) are assessed on their contribution to the achievement of a school's goals.	0.338	-0.155	**-0.400**	0.024
24	School time is structured to provide for cooperative work activity.	0.131	0.047	**0.538**	-0.057
51	Student achievement data are used to assess each teacher's performance.	0.108	-0.222	**-0.515**	-0.049
8	Staff members provide constructive feedback to each other regularly.	0.046	-0.171	0.170	**0.513**
15	Individual staff members alter their work patterns in response to feedback.	0.002	0.066	-0.126	**0.449**
59	Staff members function as a resource to each other.	0.343	0.184	0.293	**0.482**
31	Professional staff members participate on school-wide task forces and/or committees.	**0.533**	**0.453**	0.148	0.114
53	Each staff member's performance goals are reviewed with the school leadership team.	**0.489**	**-0.421**	-0.336	**0.731**
5	Data about student achievement, school services and programs are analyzed by the professional staff to aid in identifying school development goals.	0.361	0.373	-0.011	-0.022
7	The readiness level of students is considered when selecting/developing instructional programs.	-0.036	0.265	0.328	0.037

(table continued)

BEST COPY AVAILABLE

Item	Question	Factors 1	2	3	4
9	Staff development programs provide opportunities to practice newly learned skills.	0.148	0.212	0.197	0.289
11	Work groups monitor and revise their work through periodic assessment of the progress made toward goals.	0.301	0.028	-0.216	0.155
12	Instructional programs are planned cooperatively by the professional staff.	0.144	0.087	0.051	0.173
13	Staff development programs are designed to facilitate adult learning.	0.062	0.389	0.010	0.031
14	Students have input into school development goals.	-0.035	-0.230	-0.146	0.054
20	Classroom organization and activities facilitate student learning.	0.080	0.364	0.209	0.173
22	Staff members have opportunities to learn by working cooperatively with colleagues.	0.117	0.159	0.349	0.362
23	Teachers identify learning expectations for students.	0.049	0.195	0.339	0.080
25	School evaluation is a cooperatively planned system.	0.210	0.004	0.045	-0.198
27	Staff members are supervised and/or coached regularly.	0.131	0.072	0.159	0.126
28	Professional staff members are assigned to work in teams.	0.303	0.112	0.144	-0.134
29	Work groups are assessed on the extent to which work group goals are achieved.	0.249	-0.253	-0.324	-0.158
30	Students engage in cooperative learning activities.	0.015	-0.069	0.320	0.036
32	Supervision of teaching is based on cooperatively identified goals and emerging needs.	0.360	0.041	0.137	0.035
33	Students are provided with sufficient time to succeed in learning tasks.	-0.014	0.037	0.308	-0.226
39	Individual staff members are assessed on the degree to which individual performance goals are achieved.	0.252	-0.091	-0.186	-0.226
40	Staff members observe and coach each other.	0.058	-0.236	0.026	0.254
44	Individual staff members are assessed on their contribution to work group goals.	0.084	-0.024	-0.329	-0.133
48	Individual staff members are assessed on their contribution to overall school goals.	0.185	0.144	-0.206	-0.144
56	Periodic feedback from sources outside the school is used to modify work practices.	0.182	0.219	-0.009	0.021
60	Student achievement is assessed in relation to overall school goals.	0.289	0.298	0.097	-0.245

Note: Salient items had pattern/structure coefficients greater in absolute value than $|0.40|$; they appear in boldface type. The instrument items are from *School Work Culture Profile* by K.J. Snyder, 1988, Tampa, FL: School Management Institute. Copyright 1988 by K.J. Snyder. Reprinted with permission.

Factor one is titled *Continuous Improvement*. Within this factor there exists the complex interaction among goals, work structures, planning, staff development, and student success measures. What appears to be reflected is the collaborative interdependence among and within goals, staff development, program development, and student success measures. Data bases are used to establish school goals, which then guide the development of new work-structure action plans, staff development opportunities, and instruction. This tight interdependence between school planning, development, and assessment is emphasized, with a clear focus on student success measures.

In *Factor two*, the central theme is *Human Resource Development and Group Planning*. Unlike staff development practices in the past, the emphasis is on the interdependence between organizational goals and outcomes, and the function performed by training, teaching, work activity, and feedback. Goal structures in this factor are those within work units and for individual workers, which provide the context for staff development. Feedback from external and internal sources to the school generates important information to guide continuous professional improvement efforts.

Factor three centers around *Strategic Planning and Accountability*. Parents, staff, and students participate in developing the school's strategic plan, which is translated into work team and individual performance goals. Teams report progress regularly to the school's leadership where accountability is placed for improvement in the success patterns for all students. Within this factor are the instructional improvement items that center on learning strategies and their effects. This represents somewhat of a departure from traditional planning processes, which center more around leadership decision making and individual teacher implementation. Decision making and accountability have shifted, with this factor structure, to the work unit (team or department) where changes are expected in programs and services that correspond to the school's goals and to the changing needs of its student populations.

Factor four is named *Collaboration*. The common theme in the items within this factor is team work, both for professionals and for students. Time is a factor in success for both groups and suggests a developmental orientation to work. An assumption in this factor is that both students and staff members are given the necessary time to work together and to proceed. The emphasis on success corresponds to the fundamental shift to a customer focus within the quality work cultures. Continuous improvement within teams, rather than individuals and the school as a whole, is expected as students and professionals seek new kinds of outcomes.

From a conceptual and practical perspective, the second-order solution presented in Table 1 involved two central themes: (a) professional and program development (factor one) and (b) human resource development and group planning (factor two). There were also individual staff performance and collaboration clusters (factors three and four). Development emerged as the strongest theme. Planning also emerged as a strong theme. Our observations in the many schools in which we have worked in the United States reinforce these findings. The authors have seen a great institutional focus on planning and development activities. Assessment emerged as the least defined of the four work-culture areas. This finding too is very revealing. There certainly appears to be a need to establish and operationalize a set of school evaluation procedures.

Conclusion

Significantly, in the past two decades there has been a stagnation in the growth of educational productivity in America. Our study addresses that stagnation focusing on the "what and how" issues involving achievement and school organization. This article proposes a solution to the fragmentation noted positing that the lack of coherence and focus is systematic in nature. This conclusion arises from a decade of administrative involvement in American and Canadian schools and studies like the one reported in this article. This study developed the linkage of student achievement with effective school organization, offered an instrument to test the assertions, and reported the findings of the second-order analysis examining the generalized components of work culture. This type of research has been largely missing in the professional literature.

The four second-order factors extracted in this study suggest a realignment of school practices around interdependent sets of work culture features. The *Continuous Improvement* factor suggests that the purpose of schooling today has shifted from the implementation of policies and practices to responding continually to the changing needs of student populations. The focus of this factor was on development. The *Human Resource Development and Group Planning* factor reflects an alignment of school goals with training and coaching activity. Within a goal-driven context, high expectations exist for continuous improvement toward goals. The focus on this factor was on group planning. The *Strategic Planning and Accountability* factor also connects the goals and plans from all levels of the school operation and links them with expectations for meeting the changing needs of student populations. Individual staff performance emerged as the focus of this factor. *Collaboration*, the last factor, reinforces a new organizational process norm for solving problems and inventing new programs and services to meet needs.

Together these factors present a somewhat fresh picture of a school, where the focus is on improving forever the effects of programs and services on student success, where professional talent is developed continually, where strategic planning guides work toward outcomes, and where collaboration among and within groups is the norm. Perhaps this is a portrait of an educational organization for the next century.

References

Blase, J. (1993). The micropolitics of effective school-based leadership: Teacher's perspectives. *Education Administration Quarterly, 29*(2), 142-163.

Brandt, R.S. (1990-1991). On local autonomy and school effectiveness: A conversation with John Chubb. *Educational Leadership, 48*(4),57-60.

Carneval, A.P. (1989). The learning enterprise. *Training and Development Journal, 43*(2), 26-33.

Chrispeels, J.H. (1992). *Purposeful restructuring: Creating a culture for learning and achievement in elementary schools.* Washington, DC: The Falmer Press.

Chubb, J., & Moe, T. (1990). *Politics, markets and America's schools.* Washington, DC: Brookings Institute.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper and Row Publishers.

Deal, T.E. (1987). The culture of schools. In L.T. Sheive & M.B. Schoenheit (Eds.) *Leadership: Examining the elusive.* (pp. 3-15). Alexandria, VA: Association for Supervision and Curriculum Development.

Deal, T.E., & Kennedy, A.A. (1982). Culture and school performance. *Educational Leadership, 40*(5), 14-15.

Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Greene, L.E. (1991). Call it culture. *Principal, 70*(4), 4.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Hocevar, A.M. (1994). Quality standards: A guide to organizational transformation. *Wingspan, 10*(1), 6-11.

Johnson, W.L., & Snyder, K.J. (1989-1990). Planning for faculty development in America's colleges. *National Forum of Applied Educational Research Journal, 2*(1), 34-38.

Kushman, J. (1992). The organizational dynamics of teacher workplace commitment: A study of urban elementary and middle schools. *Education Administration Quarterly, 28*(1), 5-42.

Larson, C.E., & LaFasto, F.M.J. (1989). *Teamwork: What must go right/what can go wrong.* Newbury Park, CA: Sage.

Loehlin, J.C. (1992). *Latent variable models* (2nd ed.). Hillsdale, NJ: Erlbaum.

Miles, M.B., & Louis, K.S. (1990). Mustering the will and skill for change. *Educational Leadership, 47*(8), 57-61.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach.* .Hillsdale, NJ: Erlbaum.

Perkins, G. (1994). Workteams: The basic units of transformation. *Wingspan, 10*(1), 11-12.

Peters, T.J., & Austin, N.A. (1985). *A passion for excellence: The leadership difference.* New York: Random House.

Rigsby, K.L. (1994). Total quality management: A philosophy for the transformation of work cultures. *Wingspan, 10*(1), 4-5.

Rossman, G.B., Corbett, H.D., & Firestone, W.A. (1988). *Change and effectiveness in schools: A culture perspective.* Albany, NY: State University of New York Press.

SAS Institute, Inc. (1986). *SAS user's guide: statistics, statistical analysis system.* Cary, NC: Author.

Sergiovanni, T.J. (1987). *The principalship: A reflective practice perspective.* Boston: Allyn & Bacon.

Smylie, M., & Denny, J. (1990). Teacher leadership: Tension and ambiguities in organizational perspective. *Education Administration Quarterly, 26*(3), 235-259.

Snyder, K.J. (1988). *School Work Culture Profile.* Tampa, FL: School Management Institute.

Snyder, K.J., & Anderson, R.H. (1986). *Managing productive schools: Towards an ecology.* Orlando, FL: Academic Press College Division, Harcourt, Brace, Jovanovich, Inc.

Spradley, J., & McCurdy, D.W. (1996). *Conformity and Conflict* (9th ed.). New York: Addison Wesley Longman.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Stratton, A.D. (1991). *An approach to quality improvement that works.* Milwaukee: ASQC Quality Press.

Sweeney, J. (1987). Developing a strong culture. *National Forum of Educational Administration and Supervision Journal, 3*(3), 134-143.

Thompson, B., & Borrello, G.M. (1986). Second-order factor structure of the MBTI: A construct validity assessment. *Measurement and Evaluation in Counseling and Development, 18*, 148-153.

Velicer, W.F., & Jackson, D.N. (1990). Component analyses versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavior Research, 25*(1), 1-28.

Walker, B.L., & Peel, H.A. (1993). Teachers and administrators: Who's who. *Education, 114*(2), 230-232.

Appendix A
School Work Culture Productivity Model

SCHOOLWIDE PLANNING

1. *Goal Setting*: Establish annual school development goals through administrative assessment and selection and also through total staff collaborative decision making.
2. *Work Group Performance*: Designate school work groups, both teaching teams or department and task force, to which are assigned school goal objectives and action planning responsibilities.
3. *Individual Staff Performance*: Establish and operationalize a teacher performance system that includes performance standards, individual goal setting and action planning procedures, performance, monitoring, due process procedures, and evaluation.

PROFESSIONAL DEVELOPMENT

4. *Staff Development*: Develop and operationalize a school program for staff growth that emphasizes new knowledge and skills that are necessary for successful attainment of school development goals (school, work, individual).
5. *Clinical Supervision*: Develop and operationalize a peer and supervisory clinical supervision program for all teachers and teams, where

performance feedback and correctives are provided weekly.

6. *Work Group Development*: Establish a healthy work climate and develop work group skills in action planning, creative and productive group communications, problem solving, and decision making. (The competency area resulted from our research analysis).

PROGRAM DEVELOPMENT

7. *Instructional Program Development*: Establish and operationalize an instructional program that reflects up-to-date research on teaching and learning, and guides the teaching improvement efforts in the following areas: curriculum implementation, student diagnosis and placement, program planning, classroom management, teaching, and learning.
8. *Resources Development*: Facilitate staff productivity in work groups and provide necessary resources for making the school an increasingly productive unit.

SCHOOL ASSESSMENT

9. *Quality Control*: Establish and operationalize a quality control system for work groups and individuals which includes goal-based observations, conferencing, periodic progress reports and plans, and conferencing and supervisory plans.
10. *Assessment*: Establish and operationalize a set of school evaluation procedures to assess student achievement gains, teaching team and task force productivity, individual teacher performance, and total school productivity.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 45-50

Suspensions of Students With and Without Disabilities: A Comparative Study

Daniel Fasko, Deborah J. Grubb, and Jeanne S. Osborne
Morehead State University

A survey and analysis of disciplinary suspensions of disabled and non-disabled students was conducted in a rural eastern Kentucky school district based on district student records. The school district sample of 3,077 students was predominately White (98%), fairly evenly split between males and females (51% and 49%, respectively), and included 437 (14%) students with disabilities. Analysis of 213 students suspended during the 1994-95 school year indicated that no suspensions were given to minority students. There were five important findings from this study: (1) regardless of gender, students with disabilities were suspended significantly more frequently than non-disabled students, (2) regardless of school level or disability status, males were suspended significantly more frequently than females, (3) within the suspended cohort, females with disabilities were significantly under-represented relative to total numbers of suspensions tallied by gender and disability status, (4) students at middle and high school level were suspended more than elementary students, and (5) there was a total lack of minority suspensions. The implications of these findings are relevant to both practice and research and suggest the need for professional awareness, training, and development of alternate discipline strategies for students with disabilities, as well as determining the generalizability of these results.

A topic of significant concern to parents, teachers, and school administrators is misbehavior at school and resulting punishment. There has been an increase in discipline problems at schools in recent years which has escalated to a level where "students cannot learn and teachers cannot teach" (Adams, 1992, p.1). School personnel attempt to control student behavior through a variety of behavior management methods, but often resort to punitive methods of discipline such as corporal punishment, suspension, and expulsion. Although school districts have been advised for years that corporal punishment is not an effective deterrent to negative behaviors (Hart, 1987), it has taken increased litigation involving corporal punishment to cause some school districts to heed the advice of professionals in the field (and their insurance companies) to abolish the use of corporal punishment.

In the absence of corporal punishment as an option, school personnel appear to have substituted suspension to control behavior. They often express their frustration that something must be done to give students who want to learn an opportunity to learn in an environment free of disruption and disobedience (Ewashen, Harris, Porter, & Samuels, 1988). Even though most educators now recognize that external suspension is ineffective and may, in fact, be counterproductive, many school administrators view some student behaviors as too serious for in-school disciplinary options and view out-of-school suspension as the most effective disciplinary procedure (Billings & Enger, 1995; Radin, 1988).

Suspension as a form of punishment has serious educational implications in that the sanction (suspension) removes the student from the environment (school) to which he/she needs to become socially acclimated. School is not only the place where students learn the academic skills they need in order to become productive members of society, it is the place where students learn appropriate behavior, cooperation, and conformity to institutional norms (Adams, 1992). Although it would obviously be easier if schools could just teach academics, many educators believe they must take the lead in teaching students acceptable behavior (Coe, 1994).

School is, in fact, one of the more dominant cultural institutions for "socializing" our young. Suspension, on the other hand, enhances the likelihood that students do not learn the necessary social skills that enable them to become productive citizens, and actually increases students' likelihood of dropping out of school, a poor solution for all (Commission for Positive Change in Oakland Public Schools, 1992; Costenbader & Markson, 1994). In addition, students lose valuable learning time because they are not in class, and the poor achieving students who can least afford to miss school are the most

---

This is a revised version of a paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS, November, 1995. Please address correspondence to Daniel Fasko, Jr., Morehead State University, U.P.O. Box 975, Morehead, KY 40351 or at e-mail address: d.fasko@morehead-st.edu.

likely to be suspended (Commission for Positive Change in Oakland Public Schools, 1992). It is no surprise then that suspended students are typically deficient in academic skills (Neill, 1976). This is a particular problem because repeat offenders, students who are suspended more than once, have been found to account for 42 percent of all student suspensions (Costenbader & Markson, 1994). Students who have inappropriate behaviors and deficient academic skills are likely to be suspended repeatedly, increasing the likelihood they will drop out of school and never learn the behaviors or skills they need.

Student punishment is receiving increased scrutiny because several studies indicate that not all students are punished equally. For example, males are much more likely than females to be suspended, and minorities are more likely than whites to receive punishment for the same offenses, regardless of age (McFadden, Marsh, Price, & Hwang, 1992; Rossow, 1984; Sanson, Prior, Smart, & Oberklaid, 1993; Wells & Forehand, 1985; Zoccolillo, 1993). The Commission for Positive Change in Oakland Public Schools (1992) found that race is clearly a factor in many disciplinary actions and that current suspension practices violate the expectation of equal opportunities for Blacks and males in general, and for Black males in particular. Also, Black females are more likely than White females to be suspended, and White females are the least likely subgroup to be suspended from regular class for disciplinary reasons (Morgan, 1991). In addition, socioeconomic status (SES) appears to play a role in suspensions. Poor children are more likely to be suspended than average income students, particularly if there are few minorities in the school (Rossow, 1984).

In other research, age appears to be a factor in diagnosis of conduct problems, with the proportion of conduct problems increasing for older children (Wolff, 1971). In examining in-school suspensions, rates of suspension differ by building level with the number of in-school suspensions significantly greater at the high school than at the middle school level (Costenbader & Markson, 1994; McFadden, Marsh, Price, & Hwang, 1994).

Few studies were found in the literature comparing punishment of students with and without disabilities. Rose (1988) found that school personnel tend to be more tolerant of disruptive behavior and violations of school rules, but less tolerant of violent behavior, among the disabled. Rose also reported that among the disabled population, learning disabled students were most likely to be suspended, behavior disordered students were most likely to be expelled, and mentally retarded students were least likely to be either suspended or expelled. In addition, McFadden et al. (1994) found that although no significant differences in seriousness of punished misbehaviors were found between disabled and non-disabled students, the disabled students were punished at a level which exceeded their proportion in the population. On further investigation, they found that disabled students tended to demonstrate less truancy and less defiant behaviors, but exhibited more incidences of unacceptable physical contact than non-disabled students. No significant differences were found between disabled and non-disabled students in incidences and/or seriousness of fighting.

In examining in-school suspensions (ISS), Morgan (1991) found that race, gender, and disability status were all factors in suspension rates. A large percentage (18 percent) of students assigned to ISS were from special education classes. In fact, Black students from special education classes were almost three times as likely as White special education students to be placed in ISS. Of all the groups Morgan studied, females in special education were the least likely to be in ISS. Leone (1985) also found disability status related to suspension rates (i.e., behaviorally disordered teenagers not in special education are often not enrolled in school, and if they are enrolled, they are continually truant and/or suspended).

Problem

An issue of concern is how one's race, gender, school level, and disability status increase or decrease the likelihood of being suspended from school. Although the previously cited research conducted in urban areas indicates that more males get punished than do females, older students are punished more than younger, and more minorities get punished than do whites, it is appropriate to determine whether or not these findings generalize to rural school environments. Additionally, more evidence is needed to determine whether students identified as disabled are suspended at a disproportionate rate compared to non-disabled students.

Method

Sample

The population studied consisted of 3,077 students in grades 1 through 12 attending school in an eastern Kentucky school district in fall 1994. Students were predominately White (3,019, 98.1%) and fairly evenly split between males and females although males slightly outnumbered females overall (1,569 males to 1,508 females). Of 58 minority students, 36 were Black (21 males and 15 females), 12 were Asian (5 females and 7 males), 8 were Hispanic (3 females and 5 males), and the remaining 2 were American Indian/Alaskan Natives (1 female and 1 male). Four hundred thirty-seven of the 3,077 students (14.2%) were special education students referred to hereafter as 'disabled.' The disability classification of the 437 special education students was: 121 speech or language disabled, 107 learning disabled, 106 mildly mentally disabled, 45 developmentally delayed (3 through 5 year olds), 26 emotionally/behaviorally disabled, 17 functionally mentally disabled, 7 multiple disabled, 4 autistic, 2 hearing impaired, 1 deaf/blind, and 1 visually impaired.

There were 988 elementary students, 1,106 middle school students, and 983 high school students in the sample. Of the elementary students with disabilities, there were 1 Black and 77 White females and 1 Black and 168 White males. Of the 190 middle/high school students with disabilities, there were 1 Black and 56 White females, and 1 Black and 132 White males. Unfortunately, cross-tabulations of the total student population by level, race and sex were not available and the students with disabilities breakdown by level could not be partitioned further than elementary versus middle/high school students combined.

Procedure

As part of a study for the Office of Civil Rights, an extensive questionnaire of aggregate student data was completed by school district personnel for the 1994-95 school year. Data regarding the number of suspensions for misconduct during that same time period were compiled by race, gender and disability category and categorized by school level (elementary, middle, or high) as far as possible. Descriptive statistics and, where possible, Chi Square analyses were calculated to determine the results.

Results

Students Suspended

Of the 213 students suspended for behavior problems, none were minority students. Nearly 82% (175) of the suspensions were given to White males, whereas the remaining 18% (38) were given to White females. Overall, 43 of the 213 students suspended (about 20%) were categorized as disabled. Table 1 depicts percents of student suspensions by race, gender and ability group. Inspection of the table reveals first that all suspensions were given to White students; second, that males were suspended at nearly five times the rate that females were suspended; last, that the cohort of disabled students accounted for about 20% of total suspensions and 98% of suspensions to disabled students were to males.

Table 1
Percentage and Frequency of Student Suspensions
by Race, Sex, and Disability Status

	Variable	Percentage	*n*
Race			
	White	100%	213
	Black	0%	0
Gender			
	Male	82.16%	175
	Female	17.84%	38
Disabled		20.19%	43
	Male	97.67%	42
	Female	2.33%	1
Non-disabled		79.81%	170
	Male	78.20%	133
	Female	21.80%	37

Table 2 displays the frequency of disciplinary suspensions by school level (elementary, middle, high), disability cohort (disabled, non-disabled) and gender. Inspection of the table indicates that at the elementary level, 3 male students were suspended for misconduct, 2 of whom were categorized as disabled. In the middle school level, 70 males, 15 of whom were disabled, were suspended as were 14 non-disabled females. At the high school level, only one female with a disability was suspended compared to 25 males with disabilities, 23 non-disabled females, and 77 non-disabled males.

Table 2
Student Disciplinary Suspensions by Level,
Sex and Disability Cohort

School level	Disability Cohort					
	Disabled		Non-disabled		Total	
	Female	Male	Female	Male	Female	Male
Elementary	--	2	--	1	--	3
Middle	--	15	14	55	14	70
High	1	25	23	77	24	102
Total	1	42	37	133	38	175

To determine whether or not the observed differences between males and females in frequencies of disciplinary suspensions were significant, regardless of disability status (disabled or non-disabled), a Chi Square goodness of fit test was calculated. Not surprisingly, the statistic confirmed that there was a significant difference between the number of male versus female suspensions regardless of disability status throughout the school population, $\chi^2(1,N=3,077) = 87.62, p < .0001$. Suspended females were greatly under-represented in this analysis, whereas suspended males were greatly over-represented. A repetition of the same analysis using the White subset of the school population (N=3,019) as a reference, because all suspensions were to White students, yielded the same overall result (see Table 3 for both analyses).

Table 3
A Comparison of Student Suspensions by Sex:
All Students versus White Students

	All Students		White Students	
	Sex		Sex	
Status	Female	Male	Female	Male
Suspended	38	175	38	175
Not Suspended	1,470	1,394	1,440	1,366

Note. X^2 (1, N=3,077)=87.62, p<.0001 for All Students
X^2 (1, N=3,019)=87.46, p<.0001 for White Students

Two additional Chi Square analyses were calculated for this data: the first, to determine whether or not suspensions occurred equivalently across disability cohorts (disabled versus non-disabled); second, to determine whether or not, within the suspended sample, the effect of gender was equivalent across disability cohorts (disabled versus non-disabled). In the first case, a significant difference was found in disciplinary suspensions across disability cohorts, $\chi^2(1,N=3,077) = 6.21, p = .01$, in the total student population. Suspensions of students with disabilities were over-represented relative to the size of that subgroup. Again, repetition of the same analysis using the White subset of the student population (N=3,019) as a reference, yielded the same overall result. Finally, it was determined that the effect of gender was not equivalent across disability cohorts within the suspended students group, $\chi^2(1,N=213) = 7.57, p=.006$. Females with disabilities were significantly under-represented relative to the total number of females suspended and the total number of students with disabilities suspended (see Tables 4 and 5 for both analyses). Due to limitations of the system data set, no further analyses were considered appropriate.

Table 4
A Comparison of Student Suspensions by Disability Status:
All Students versus White Students

	All Students		White Students	
Status	Disabled	Not Disabled	Disabled	Not Disabled
Suspended	43	170	43	170
Not Suspended	394	2,470	390	2,416

Note. X^2 (1, N=3,077)=6.21, p=.01 for All Students
X^2 (1, N=3,019)=5.87, p=.02 for White Students

Table 5
Student Suspensions by Disability Status and Sex

Sex	Disabled	Not Disabled
Female	1	37
Male	42	133

Note. X^2(1,N=213)=7.57, p=.006

Discussion

Contrary to previous studies, the results from this small study indicate that more Whites were suspended than is proportional for the school population because no minorities received suspensions. Given the nature of the population in the school district, with an extremely small number of minorities (just over 2 percent) who are predominantly the children of university faculty members and university students, this result was not surprising. Additionally, the data support previous research: (1) that disabled students are over-represented in suspensions (20% of the suspensions were given to students categorized as disabled when only 14% of the student population was so classified); (2) that males were punished more than were females in the schools, given that almost 5 times as many suspensions were to males than to females (82% compared to 17%); (3) that the least likely subgroup to be suspended was female categorized as disabled (only one suspended student in the present sample was a female with disabilities); and (4) students at the middle and high school levels were suspended appreciably more often than elementary school children and there was a complete lack of minority student suspensions. In general, findings support the results of research conducted on suspension in urban schools.

As might be expected, the majority of school suspensions occurred in the middle and high schools, which supports McFadden et al's. (1992) results on increased incidence of suspension as age and grade level increase. Perhaps, more students in the elementary grades adhere to the code of conduct in their respective schools than do students in middle or high school or perhaps the public is less tolerant of suspensions of elementary school age children. Alternatively, elementary student misconduct may not be perceived to be as threatening or serious to adults in authority as is misconduct in older students. If there is, in fact, a difference in severity of behavioral problems in adolescents, it may be due to factors such as peer or environmental influences, sex differences, and emotional and behavioral difficulties as a result of puberty (Fry & Gabriel, 1994; Sanson et al., 1993; Zoccolillo, 1993). It is also possible that the difference in suspensions between elementary, middle, and high school might be due to differences in school climate, school philosophy, or discipline practices. For example, the Commission for Positive Change in Oakland Public Schools (1992) found that suspension works as a short term release valve for a school and serves a function for the school, not the student. Similarly, Rossow (1984) found that whether or not a student is suspended is more a matter of the behavior differences between administrators than differences in behaviors between students. Several studies have indicated that suspension is not just due to better or worse student behavior, but may be due more to school climate, discipline policies, and administrator/teacher behavioral differences (Commission for Positive Change in Oakland Public Schools, 1992; Costenbader & Markson, 1994; McMahon & Wells, 1989; Rossow, 1984). In fact, Rossow (1984) found that it is mostly the discretion of the principal that determines what happens to an individual student and having discretion increases the opportunity for subtle prejudices.

If the Commission for Positive Change in Oakland Public Schools (1992) is correct in its supposition that suspension serves as a short term release valve for a school, those students whose behavior is relatively worse would be targeted for the pressure release. In any group, behavior can be considered relative to the group instead of in terms of an absolute standard, so there will always be a group potentially targeted as the troublemakers who prevent all of the rest of the students from enjoying the educational atmosphere to which they are entitled. If the teachers and administrators expect that there will be a "few" behavior problems so severe that the students need to be removed from school, then the worst relative behavior problems will be the ones targeted for suspension. Interestingly, Morgan (1991) suggested that students who are suspended may be operating under a self-fulfilling prophecy and engaging in behavior that is expected of them, even if what is expected is misbehavior.

Because students need to be in school to gain high quality academic and social educations, schools should explore in-school disciplinary alternatives to suspension. Intervention strategies, such as conflict resolution and other behavioral shaping techniques, are based on the philosophy that children should be in school where they can learn appropriate behaviors. It has been found that behavior training programs can significantly improve student behavior (McMahon & Wells, 1989), whereas, punishments, such as suspension, remove the students from the learning environment and inform students what they should not do.

Punishment alone, and suspension in particular, does *not* change the behavior the school found to warrant the suspension and does *not* teach a student acceptable behavior (Commission for Positive Change in Oakland Public Schools, 1992; Hartwig & Ruesch, 1994). Coe (1994) insists that the fundamental issue is for schools to promote positive student behaviors and to discourage inappropriate student behaviors by making students accountable for their actions. The Commission for Positive Change in Oakland Public Schools (1992) determined that academic success is the key to good behavior. They found that behavior problems and discipline referrals are reduced when student achievement is improved and recommended that schools focus on good teaching and learning to maintain good student behavior. An additional reason for implementing intervention strategies is that children who exhibit severe behavior problems and do not receive treatment have an increased likelihood of engaging in later criminal behavior (Loeber, 1982). It sounds simplistic to suggest that improved teaching and learning could reduce behavior problems, however, that is often what is needed.

Perhaps also, males with disabilities are overrepresented in suspensions because school districts are suspending them rather than providing appropriate educational programming. Further research is needed to determine why this overrepresentation of suspended males with disabilities exists.

In summary, future research should continue to be directed toward determining the effectiveness of proactive interventions in reducing the number and proportion of suspensions in groups at high risk for punishment; that is, males, minorities, adolescents, and special education students. Lastly, research should look at the efficacy of interventions and punishments in rural vs urban settings. It remains to be determined whether males, older students, and students with disabilities actually engage in more punishable behavior or whether students are punished differently for the same acts. Also, more research is needed comparing the nature of the behavioral problems exhibited by the different groups.

References

Adams, T. (1992). *Public high schools: The uses of rehabilitative and punitive forms of discipline: A final report.* Washington, D.C.: Office of Educational Research and Improvement.

Billings, W. H., & Enger, J. M. (1995, November). *Perceptions of Missouri high school principals regarding the effectiveness of in-school suspension as a disciplinary procedure.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Coe, D. (1994). We must act now to make our schools safe. *NASSP Bulletin, 78,* 108-110.

Commission for Positive Change in Oakland Public Schools, CA. (1992). *Keeping children in school: Sounding the alarm on suspensions.* Oakland, CA: Urban Strategies Council, 672 13th Street, Suite 200, Oakland, CA 94612. (ERIC Document Reproduction Service No. ED 350 680)

Costenbader, V. K., & Markson, S. (1994). School suspension: A survey of current policies and practices. *NASSP Bulletin, 78,* 103-107.

Ewashen Jr., G., Harris, S., Porter, D., & Samuels, K. (1988). School suspension alternatives. *Education Canada, 28,* 4-9.

Fry, P., & Gabriel, H. (1994). Preface: The cultural construction of gender and aggression. *Sex Roles, 30,* 165-167.

Hart, S. N. (1987). Psychological maltreatment in the schools. *School Psychology Review, 16,* 169-178.

Hartwig, P., & Ruesch, M. (1994). *Disciplining students with disabilities: A synthesis of critical and emerging issues.* Alexandria, VA.: National Association of State Directors of Special Education.

Leone, P. E. (1985). Suspension and expulsion of handicapped pupils. *The Journal of Special Education, 19,* 111-121.

Loeber, R. (1982). The stability of antisocial and delinquent child behavior: A review. *Child Development, 53,* 1431-1446.

McFadden, A. C., Marsh, G. E., Price, B. J., & Hwang, Y. (1992). A study of race and gender bias in the punishment of school children. *Education and Treatment of Children, 15,* 140-146.

McFadden, A. C., Marsh, G. E., Price, B. J., & Hwang, Y. (1994). A study of race and gender bias in the punishment of handicapped school children. *The Urban Review, 24,* 239-251.

McMahon, R. J., & Wells, K. C. (1989). Conduct disorders. In E. J. Mash & R. A. Barkley (Eds.) *Treatment of Childhood Disorders* (pp. 73-132). New York: Guilford.

Morgan, H. (1991, April). *Race and gender issues: In school suspension.* Paper presented at the Annual Meeting of the American Education Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 028 922)

Neill, S. B. (1976). *Suspension and expulsion: Current trends in school policies and programs.* Arlington, VA.: National School Public Relations Association.

Radin, N. (1988). Alternatives to suspension and corporal punishment. *Urban Education, 22,* 476-495.

Rose, T. L. (1988). Current disciplinary practices with handicapped students: Suspensions and expulsions. *Exceptional Children, 55,* 230-239.

Rossow, L. F. (1984). Administrative discretion and student suspension: A lion in waiting. *Journal of Law & Education, 13,* 417-440.

Sanson, A., Prior, M., Smart, D., & Oberklaid, F. (1993) Gender differences in aggression in childhood: Implications for a peaceful world. *Australian Psychologist, 28,* 86-92.

Wells, K. C., & Forehand, R. (1985). Conduct and oppositional disorders. In P. H. Bornstein & A. Kazdin (Eds.), *Handbook of clinical behavior therapy with children* (pp. 218-265). Homewood, IL: Dorsey.

Wolff, S. (1971). Dimensions and clusters of symptoms in disturbed children. *British Journal of Psychiatry, 118,* 421-427.

Zoccolillo, M. (1993). Gender and development of conduct disorder. *Development and Psychopathology, 5* (1-2), 65-78.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 51-55

The Effects of Specific Interventions on Preservice Teachers' Scores on the National Teacher Exam

Hunter Downing, Sue Austin, and Eileen Lacour
Southeastern Louisiana University

Nancy Martin
University of Texas, San Antonio

Students preparing to take the National Teachers' Exam for the first time were randomly assigned to one of four groups (test anxiety reduction, learning/test taking strategy, domain specific knowledge, and control) prior to taking the exam. The groups were composed of volunteers from teacher education classes who attended two 2-hour sessions designed to increase the specific learning strategy tactics, their knowledge of content areas covered on the NTE, or to reduce test anxiety. Scores on the NTE were used as the dependent measure in analyzing the results. Students who received training in learning test-taking strategies performed significantly better than students in other groups. Implications for students who take the NTE are discussed.

The public and media often use standardized tests to measure the effectiveness of education. During the 1980s, a variety of teacher testing programs were legislated throughout the United States (Popham, 1990). Although controversial, some form of teacher testing has been mandated in at least 35 states; of those, 17 states require the National Teacher Exam (NTE) (Egan & Ferre, 1989).

The long-term goal of teacher testing programs is to improve the quality of the teacher workforce. To that end, teacher exams are used for a variety of purposes: as a screening device for students who wish to enter a teacher education program, as a pre-certification requirement, or as a requirement for re-certification of experienced teachers (Egan & Ferre, 1989). Considering the high stakes associated with these test results, identifying successful test preparation methods is of interest to future and current educators as well as the general public. Michael and Edwards (1991) explain, "If students were sophisticated test takers, the scores would be more valid representations of what students knew, because types of test questions or poor answering strategies would not decrease students' scores" (p. 106).

Ethical means of improving standardized test scores have been of interest for more than 40 years (e.g., Dyer, 1953, 1987; Michael & Edwards, 1991). Within this corpus of research, subject pools tend to focus on elementary and secondary levels, but university participants are also represented. Across the spectrum, results tend to conclude that coaching can be helpful in producing improved measurement outcomes (e.g., Alcorn, 1990; Chicago Public Schools, 1987; Deaton, Halpin, & Alford, 1987; Ornstein, 1993; Thomas, 1986). The literature addresses three general aspects of test preparation: improving test-wiseness (i.e., test-taking skills and strategies), focusing on content and ability areas measured by the specific test, and lessening test anxiety by instructing test takers regarding format, directions, and time management (e.g., Dyer, 1987; Michael & Edwards, 1991; Ornstein, 1993).

In an examination of subjects who had to retake one part of the NTE core battery, Alcorn (1990) focuses on test-related actions taken by students between their initial failure and subsequent success. Results suggest that while the Communication Skills Test was the portion most often failed, subjects relied heavily on special preparation which included the use of a commercial study guide and attending workshops. Additionally, review of course materials and tutoring in specific content areas were found to be helpful to these students (Alcorn, 1990).

Current research in the field of metacognition suggests that the theoretical and empirical basis of cognitive strategies instruction is closely related to the major principles of constructivism (Harris & Pressley, 1991). Harris and Pressley present cognitive instruction models in the areas of reading comprehension, written language,

---

Hunter Downing and Sue Austin are assistant professors of Counselor Education at Southeastern Louisiana University. Eileen Lacour, now deceased, was an assistant professor in the College of Basic Studies at the same institution. Nancy Martin is an assistant professor of Educational Psychology at the University of Texas, San Antonio. Correspondence regarding the paper should be addressed to Hunter Downing, Southeastern Louisiana University, Dept. of Counseling, Family Services, Educational Leadership, 305 SLU, Hammond, LA 70402 or by e-mail to hdowning@selu.edu.

and memory. Training in the use of cognitive techniques such as problem solving and considering all alternatives may enhance individuals' abilities to perform well in evaluative settings and improve performance on standardized examinations (e.g., Anastasi, 1988; Frierson, 1984). Results are particularly encouraging among learning disabled and minority students (Scruggs, 1986).

A basic premise of the research on test anxiety is that students experiencing test-anxiety do not perform to their potential, resulting in an underestimation of their abilities by educators (DuBois, 1987). Hill and Wigfield (1984) describe school intervention studies where new evaluation procedures and teaching programs have been successfully developed to help students perform better in assessment situations. DuBois (1987) reviews a variety of intervention techniques including adaptations of testing situations, learning strategies, test-wiseness instruction, and coping skills training which appear effective in alleviating test anxiety. Systematic desensitization has been used to successfully treat many types of anxiety or fear, including test anxiety (e.g., Austin, Partridge, Wadlington, & Bitner, 1995; Kosta & Galassi, 1974). This method is useful when a person has the ability to handle the situation but avoids the situation or does poorly in it due to anxiety (Cormier & Cormier, 1991).

This study was designed to determine which of three interventions would yield the most significant effect on NTE scores. Specifically, the following four objectives were formulated:

1. To evaluate the effectiveness of learning strategies training on the participants' success on the NTE.
2. To determine if instruction in content specific knowledge improves NTE scores.
3. To distinguish which intervention facilitates a higher success rate for preservice teachers on the NTE.
4. To ascertain if training in systematic desensitization skills to reduce test anxiety results in higher scores on the NTE than other specific interventions.

Methodology

Participants

Participants were 49 students enrolled in undergraduate education courses at a mid-sized southern university who signed up for a free NTE seminar. There were approximately 200 students who were enrolled in classes that were invited to attend the seminar. The mean age of the participants was 23.34 ($SD = 1.22$), and included 43 women and 6 men. There were 3 African-American participants, 2 Hispanic participants, and the rest were Caucasian.

Procedure and Instruments

Students who registered for the seminar were randomly assigned to one of four groups. The four treatment groups were test-anxiety reduction, test-taking strategies, domain-specific knowledge, and control. The students signed a consent form acknowledging that the purpose of the seminar was to inform them about the NTE, and that it did not ensure that participation would improve their scores.

Two weeks prior to the NTE exam date, all groups met with different instructors for two 2-hour sessions. Only students who attended both sessions were included in the study.

All four groups were given general information about the NTE, such as where the test was to be administered, what to bring with them to the test site, and the general format of the exam. Following that, each instructor presented the specific interventions to each of the groups.

Subjects in the test-anxiety reduction group participated in two 2-hour sessions involving systematic desensitization. During the first session, the students were asked to collectively determine ten test-related situations which cause them anxiety. The ten items were written on individual index cards. Subjects were then asked to rank them in order from least anxiety-provoking to most anxiety-provoking. Once the test anxiety hierarchy was established, subjects were taught deep muscle relaxation techniques as described by Cormier and Cormier (1991), and they were told to practice these techniques daily at home. In addition, subjects were told to think of a special place or situation which they considered extremely relaxing. During the actual group sessions, subjects were asked to imagine themselves in the test-anxiety producing situations while maintaining a relaxed state. They were to signal the instructor when they began to feel tense or anxious. At the first sign of anxiety the participants were told to imagine themselves in their special place until they were again completely relaxed. The instructor then took them back to a test situation that was less anxiety-producing and gradually progressed through the hierarchy. This process was followed for both two-hour sessions. Students were also given a handout of the relaxation procedure, informed of ways of using relaxation strategies on the day of the test and encouraged to practice the exercises that they had learned in the days preceding the exam.

The test-taking-strategies group received general instruction in metacognitive and test-taking strategies. For example, students were given a handout with sample questions from NTE practice tests, and allowed to go through the items as the instructor pointed out relevant tips, such as spotting tricky words including *not* and *except.* Other topics covered were *Knowing when to guess, Using time effectively*, and *How to concentrate on the test.* Instruction in general cognitive skills as suggested by Anastasi (1988), such as carefully analyzing problems, avoiding impulsive answers, and considering all alternatives, was an important component of this intervention. Students were encouraged to review the test format the night before the exam, and to go to the exam well-rested.

The domain specific group was given specific instruction in subject matter found on the General Knowledge portion of the NTE. Approximately one hour was spent on each of the following areas: Mathematics, Social Studies, Science, and Art & Literature. Test preparation materials from the Educational Testing Service were used for instruction with this group. Overhead transparencies were prepared from practice tests obtained from ETS, and this material was supplemented by the instructor and with a video-tape series on NTE preparation. The participants were encouraged to go over their notes at least twice before taking the exam.

The control group received instruction on the content of the Professional Knowledge portion of the exam. Because scores on this part of the exam were not utilized in the study, this group was used as the control.

At the beginning of the first session, all students filled out a Test Anxiety Inventory (TAI). Other demographic information was also collected, such as age, gender, and college classification. A one-way analysis of variance (ANOVA) on the TAI was performed in order to ensure that there was no significant difference between groups on this measure.

After the results from the NTE were reported, analyses were conducted to ascertain whether any of the interventions appeared to be helpful in improving scores on the Communication Skills or General Knowledge portions of the NTE.

Results

A multivariate analysis of variance (MANOVA) was used to determine differences between the four groups on the dependent variables, which were the Communication Skills scores and the scores on General Knowledge. Means and standard deviations for all groups are shown in Table 1.

Table 1
Means and Standard Deviations for NTE scores

	General Knowledge Scores	
	Means	SD
Domain-specific group	652.58	8.73
Test-taking strategies group	656.50	6.33
Relaxation group	650.73	6.81
Control group	647.50	6.30
	Communication Skills Scores	
	Means	SD
Domain-specific group	655.50	7.24
Test-taking strategies group	658.24	6.85
Relaxation group	655.86	6.13
Control group	653.00	7.81

Results of the MANOVA indicated that there was a significant difference between groups on the General Knowledge scores, $F(4,54) = 3.09$, $p < .05$. There was no significant difference in scores on the Communication Skills scores. Post hoc analyses using Tukey's multiple range test revealed that the test-taking strategies group had significantly higher scores than the control group on General Knowledge. There were no other statistically significant differences found. Table 2 presents the results of the MANOVA.

Discussion

Standardized tests are used for a variety of reasons within the teaching profession, including screening prospective teacher education students, pre-certifying beginning teachers, and re-certifying experienced teachers (Egan & Ferre, 1989). Because standardized tests are used as a means to determine an individual's future, it is imperative that teacher education programs make every effort to identify and implement methods to assist students in succeeding at this often frightening proposition.

Table 2
Multivariate Analysis of Variance (df: 4,54)

Variable	SS	MS	F	Sig. of F
Gen. Kn.	464.65	154.88	3.09	.036*
Com. Sk.	149.78	49.92	.58	.631

*p< .05

Current literature suggests several methods which have been considered useful in improving test scores. Methods have tended to focus on coaching (Alcorn, 1990; Chicago Public Schools, 1987; Deaton, Halpin & Alford, 1987; Ornstein, 1993; Thomas, 1996); improving test-wiseness (DuBois, 1987), improving knowledge in content and skills areas (Alcorn, 1990; Gilli & Gilli, 1987); reducing test anxiety through instructions regarding format, directions and time management (Dyer, 1987; Michael & Edwards, 1991; Ornstein, 1993); cognitive instruction models in the areas of reading comprehension, written language and memory (Harris & Pressley, 1991); cognitive problem solving techniques (Anastasi, 1988; Frierson, 1984); and systematic desensitization (e.g., Austin et al., 1995).

The purpose of this study was to determine whether one or all of several methods used with other testing situations would be successful in improving NTE scores of undergraduate teacher education majors. Methods included systematic desensitization, learning strategies training, and instruction in content-specific knowledge. Not only did statistical analyses suggest positive results, but also feedback from participants indicated a positive reaction to the assistance they received.

Since all of these participants were volunteers, it is possible that their willingness to attend orientation sessions affected their scores, so caution must be used in generalizing the results. Also, because of the low N, results could be sample specific. However, the primary finding of this study is that instruction in test-taking strategies led to significantly higher scores on the General Knowledge portion of the NTE. While the other two interventions, instruction in domain-specific knowledge and systematic desensitization, also elicited higher scores than the control group, they were not significantly higher. This is surprising considering previous research which has found significant positive results with these two methods in the treatment of test anxiety (Austin et al., 1995; Alcorn, 1990; Chicago Public Schools, 1987; Deaton et al., 1987; Ornstein, 1993; Thomas, 1996).

It may be that the students in the domain-specific and the test-anxiety groups did not do as well as the test-taking-strategies group because of the limited exposure to the treatments. The topics in the domain specific group were covered in only 4 hours, and it was not possible to go into depth in any of the areas. In addition, students may have varied on their prior knowledge, as well as their ability levels which could have influenced the effect of the training. It would probably be more helpful to determine students' current knowledge in the content area prior to receiving extended instruction in their specific areas of perceived weakness in order for there to be significant gains in scores. Since the domain-specific group did achieve higher scores than the control group, it would seem that additional instruction in the content area would be useful. Continuing education workshops in content areas might be useful to those registering for the NTE. In addition, many universities require remediation in math and reading to students scoring below a specific cut-off on the ACT. Similar programs could be added in the areas of social studies and science, which might enhance NTE scores.

Similarly, the test-anxiety group had little time to learn and practice the skill of progressive relaxation. Although the participants were urged to continue to practice daily until the day of the exam, it was not possible to ascertain how many students actually followed this suggestion. To improve the result of systematic desensitization, longer-term treatment with the addition of verbal therapy might be more beneficial. Providing individual counseling for all test-anxious students is unrealistic; however, other methods of teaching the same goal are available, such as classroom instruction or orientation workshops.

Test-taking strategies are more easily taught, and more readily applied than the other two interventions. For this group of students, instruction in general meta-cognitive strategies apparently helped them earn higher scores on the NTE.

Because it is often difficult to entice students who are first-time takers of the NTE to attend orientation sessions, it is important to spend the instructional time effectively. It would seem that concentrating on helping students learn various test-taking strategies would be helpful if there is a limited amount of time available for instruction before the exam date.

Future research should attempt to obtain a larger number of participants for the interventions. Also, there should be more sessions for the interventions, so that participants in the domain-specific and relaxation groups have a chance to gain more knowledge and skills. Finally, combining two treatment groups would, perhaps, lead to a greater improvement in scores.

References

Alcorn, B. (1990). *Retaking the NTE core battery: What helps*? ERIC Document No. ED 334 255.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Austin, S., Partridge, E., Wadlington, E., & Bitner, J. (1995). Prevent school failure: Treat test anxiety. *Preventing School Failure, 40*, 10-13.

Chicago Public Schools (1987). *A comparison of the effectiveness of four test preparation programs*. Final Evaluation Paper. ERIC Document No ED 318 739.

Cormier, W., & Cormier, L. S. (1991). *Interviewing strategies for helpers* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Deaton, W. L., Halpin, G., & Alford, T. (1987). Coaching effects on California Achievement Test scores in elementary grades. *Journal of Educational Research, 80*, 149-155.

DuBois, J. (1987). *Recognition and intervention of interfering test anxiety: An annotated bibliography*. ERIC Document No. ED 292 815.

Dyer, H. S. (1953). Does coaching help? *College Board Review, 19*, 331-335.

Dyer, H. S. (1987). The effects of coaching for scholastic aptitude. *NAASP Bulletin, 71*, 46-50, 52-53.

Egan, P. J., & Ferre, V. A. (1989). Predicting performance on the National Teacher Examinations Core Battery. *Journal of Educational Research, 82*, 227-230.

Frierson, H. (1984). *Effects of test-taking instruction on a health professional certifying examinations: An evaluation*. ERIC Document No. ED 246 094.

Gilli, L. M., & Gilli, A. C. (1987). Preparing for the National Teachers Exam. *Vocational Education Journal, 62*(6), 24-26.

Harris, K, & Pressley, M. (1991). The nature of cognitive strategy instruction interactive strategy construction. *Exceptional Children, 57*(5), 392-404.

Hill, K., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *Elementary School Journal, 85*(1), 106-126.

Kosta, M. P., & Galassi, J. P. (1974). Group systematic desensitization versus covert positive reinforcement in the reduction of test anxiety. *Journal of Counseling Psychology, 21*, 464-468.

Michael, N, & Edwards, P. A. (1991). Test preparation programs: Counselors' views and involvement. *The School Counselor, 39*, 98-106.

Ornstein, A. C. (1993). Coaching, testing, and college admission scores. *NAASP Bulletin, 77*, 12-19.

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Scruggs, T. (1986). *The administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school students: Year three final report*. ERIC Document No. ED 311 652.

Thomas, M. C. (1986). *Effective methods for increasing scores on standardized tests*. ERIC Document No. ED 272 811.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 57-66

Assistant Principals' Concerns About Their Roles in the Inclusion Process

Louise L. MacKay
East Tennessee State University

Patricia D. Burgess
University of Kentucky

This study surveyed 62 assistant principals from schools in North Carolina, Tennessee, and Virginia to assess their level of concern about their skills in the areas of visionary leadership, collaboration and supports for staff and students. The researchers narrowed the focus of the Concerns Questionnaire for Change Facilitators to reflect findings of the National Survey on Inclusive Education. Based on frequency counts and descriptive statistics of this study, results suggested that the three areas necessary for successful inclusion by the national survey were also areas of concern about roles and responsibilities of assistant principals.

Rationale and Parameters of Inclusion

The nationwide movement to implement successful inclusion programs is an attempt to educate children with special needs to the maximum extent appropriate in the school and classroom they would otherwise attend (Rogers, 1993). The paradigm shift to inclusion requires leadership commitment and support with professional development and assistance to prepare general educators and special educators to incorporate students with disabling conditions into the general education classrooms. Program success requires joint planning, collaboration, and leadership. "Countless studies have demonstrated that innovation without supportive consultation on an ongoing basis does not have lasting results" (MacKay, 1994, p. 6). This study sought to identify the leadership concerns of assistant principals in visionary leadership needs in North Carolina, Tennessee, and Virginia, and to assist them in implementing and maintaining successful inclusion programs.

Background Information

The combination of special and general education governance and forms of service delivery produced the concept of inclusion. The National Center on Educational Restructuring and Inclusion (NCERI) defined inclusion as:

> Providing to all students, including those with severe handicaps, equitable opportunities to receive effective educational services, with the needed supplementary aids and support services, in age appropriate classes in the neighborhood schools, in order to prepare students for productive lives as full members of the society. (NCERI, 1994, p. 4)

The basic premise of inclusion is to provide an appropriate education for students with special needs by combining the content knowledge of the general educator and the exceptionality knowledge of the special educator.

The passage of the Individuals with Disabilities Education Act (IDEA) sparked an increased awareness for better public education for individuals with disabling conditions. IDEA guaranteed individuals with special needs the right to: employment, independent living, participation in community activities, and free access (Pues, 1990). IDEA reflected and supported the inclusion concept of combining special education delivery with general education delivery to help children with disabilities to perform as much like nondisabled students as possible (Fuchs & Fuchs, 1995). Additionally, inclusion promotes the concept that public schools should realistically prepare all students for after school and post school experiences and should provide the skills and knowledge to contribute to a unified society (Stainback & Stainback, 1985).

Louise L. MacKay is an associate professor in the department of Educational Leadership and Policy Analysis at East Tennessee State University. Patricia D. Burgess is a technical assistance provider at the MidSouth Regional Resource Center at the University of Kentucky. Please direct all correspondence to Dr. Louise L. MacKay, Associate Professor, East Tennessee State University, Department of Educational Leadership and Policy Analysis, Box 70550, Johnson City, TN 37614, (423) 929-6716. Internet: MacKay@ACCESS.ETSU.TN.EDU

Joseph Fisher, assistant commissioner of special education for the Tennessee State Department of Education, stated:

> The recognition of inclusion by the federal government and advocates of children with disabilities is consistent with IDEA and its companion Code of Federal Regulations in that inclusionary educational practices are meant to involve individualized education programs using appropriate supplementary aids and services based on the needs of each child. (Personal communication, December 7, 1993)

Fisher (personal communication, December 7, 1993) continued his support for inclusion by making the following points:

1. Proper support and training should be provided to all who will be involved in the planning of the inclusionary educational practice.
2. Proper support and training must also be provided to the students served in inclusive settings.
3. School systems should not expect efforts which encourage services in the least restrictive environment to be less costly or require less effort than traditional special education services.
4. The continuum of services must be made available as needed for all children. The common thread that permeates the concept of inclusion is the necessity of individualization based on the needs of each student.

Supporting Structure for Implementing the Inclusion Model

The inclusion transformation requires the acquisition of new knowledge and more emphasis on visionary leadership. Cunningham and Gresso (1993) stated that educational leaders must be provided an opportunity to develop new knowledge, skills, and abilities related to new ideas and directions. They further asserted that "adults learn as a result of their own personal and professional needs, and no developmental activity will be successful unless the need is recognized by the individual" (1993, p. 189). The National Association of State Boards of Education (NASBE, 1994) researched the training and development needs of teachers involved with inclusive programs. Teachers identified the following needs: (a) provide problem solving time through staff training, (b) provide training in instructional methods and teaching strategies, and (c) provide training on implementation of change. These findings were supported by the national survey conducted by the National Center on Educational Restructuring and Inclusion (NCERI), The Graduate School and University Center, and The City University of New York to identify successful inclusive education programs and factors necessary for inclusion to succeed. Essential factors included: (a) visionary leadership, (b) collaboration, (c) supports for staff and students, (d) refocused use of assessment, (e) funding, (f) effective parental involvement, and (g) exposure to models and classrooms practices that support inclusion (NCERI, 1994).

According to Katsiyannis, Conderman, and Franks (1996), educational leaders "are instrumental in and ultimately responsible for providing the necessary leadership . . . and fostering collaboration of general and special education teachers by defining roles, responsibilities and processes for program delivery" (p. 82). Additionally, principals are instrumental in providing necessary supports to staff and students. These supports may include "providing ongoing staff development opportunities, providing release time for team planning and preparation activities, establishing coaching systems to maintain and reinforce instructional skills, . . . and evaluating inclusionary programming outcomes" (Katsiyannis, Conderman, & Franks, 1996, p. 82).

According to Murphy and Schiller (1996), school principals who are most successful in fundamental reform endeavors often perform five functions: "(1) support the development of a shared vision, (2) create a network of supportive relationships, (3) allocate adequate resources to facilitate vision, (4) keep the staff informed, and (5) promote teacher development" (p. 31). If the vision of an appropriate education for every child is to become a reality in our diverse and rapidly changing society, creative ways of utilizing staff and resources are imperative. Lombardi (1994) addressed this point when he stated that "effective inclusion will be impossible to achieve without the support of school administrators. For example, principals must be able to identify teachers who will be successful on inclusion teams. They must allow time for team planning and problem solving" (p. 24).

Villa and Thousand (1995) support these role changes and training needs. They pointed out administrators not only need to provide time for team planning, but they should participate as members of teams "that invent solutions to barriers inhibiting the successful inclusion and education of a child" (Villa & Thousand, 1995, p. 69). They further pointed out that administrators should lead the development of a school vision and mission that include special education students in general education classes and then articulate this vision to the teachers, students, parents, and public.

Fullan (1996) described vision building as the connection between moral purpose and the forces of change. He suggested that educators, teachers and administrators are the moral change agents in society and that purposeful change is the new norm in education. Goodlad (1990) described the basic rationale for teaching in post-modern society by stating that "teaching in schools carries with it a moral imperative which includes developing educated persons who acquire an understanding of truth, beauty, and justice against which to judge their own and society's virtues and imperfections" (pp. 48-49). The vision of an appropriate education for each child is a vision that must beckon a successful inclusive school.

For too long, the inclusion of all children in general school programs has been perceived as an additional burden on an already overburdened educational system. One way this can be ameliorated is to utilize more creatively the personnel that exist in the school. Clearly, school administrators need to participate in and be supportive of the vision of education for all children. However, traditional ways of thinking about role responsibilities have seldom been successful in bringing about sustained change. Special education populations and approaches are often a mystery to administrators, and inclusion has only complicated an inadequate area of training. For instance, Keaster (1996) found that 51% of Louisiana principals had no training of any kind regarding special education and 49% had never had a special education course. In our discussions with assistant principals who participated in this study, very few of them had any training or support for the process of working with diverse and special needs students. Much of their resistance to inclusion seemed to be based on their feelings of inadequacy as to what decisions should be made for the best interests of all concerned.

Statement of the Problem

To prepare, implement, and maintain successful inclusion programs, relevant and appropriate visionary leadership, collaboration, and supports for staff and students are necessary. Assistant principals are often assigned the role of placement coordinators for students with disabilities and therefore, need extensive development in the areas of visionary leadership, collaboration, and providing supports for students and staff.

Method

Sample

Sixty-two assistant principals participated in this study as part of the Assistant Principal's Academy in Abingdon, Virginia. These assistant principals were selected by their school systems to participate in the Assistant Principal's Academy to address what they had indicated as major concerns related to their roles as assistant principals.

The Assistant Principal's Academy was a four-phase leadership development program designed and implemented as a collaborative effort of the educational leadership departments of Appalachian State University, East Tennessee State University, and Virginia Polytechnical Institute. The Academy addressed a wide variety of needs or concerns identified by the participating assistant principals through a needs assessment. One full day session addressed the area of inclusion of students with disabilities in general education environments. The assistant principals completed the Concerns Questionnaire for Change Facilitators as an introduction to the inclusion session.

Instrumentation

The Concerns Questionnaire for Change Facilitators, developed by Hall, Newlove, George, Rutherford, and Hord (1991) was used for this study. The structured survey questions were designed to determine what educational leaders "are thinking about regarding [their] responsibilities as change facilitators" (Hall, Newlove, George, Rutherford, & Hord, 1991, p. 56). The format of the questionnaire responses was a Likert-type scale. Participants were asked to indicate the degree of concern they felt within their present roles. The eight point scale ranging from "irrelevant" (0) to "very true of me now" (7) was compressed into a five-point scale for recording purposes.

To establish logical validity, or content validity, the researchers "assumed the role of 'experts' and determined whether the test or test items were content valid for the study" (Gay, 1996, p. 140). Survey questions were logically combined into areas of visionary leadership, collaboration, and supports for staff and students. Frequency counts for each subsection are recorded in Table 1.

Research Design

This study involved descriptive statistics research methods to clarify and summarize numerical data generated by assistant principals to determine the perceptions about their comfort at being change facilitators. Descriptive statistics to determine the mean and standard deviation for each item were generated. The mean generated a sense of the middle or average score for a variable, while variability, or standard deviation, was used with the mean to show how the other scores are distributed around the mean (Hittleman & Simon, 1992). Frequencies were calculated to generate frequency count recordings of each question and compilation of subsections. Frequency count recording was useful to determine the participants' present perceptions of their roles as change facilitators.

Table 1
Percents of Frequency Counts for Concerns Questionnaire for Change Facilitator

Visionary Leadership Survey Items	Percent
I would like more information about the purpose of "Inclusion"	
Irrelevant	1.6
Not true of me now	6.4
Somewhat true of me now	33.9
True of me now	17.7
Very true of me now	40.3
I am more concerned about facilitating use of "Inclusion"	
Irrelevant	0.0
Not true of me now	11.3
Somewhat true of me now	25.9
True of me now	21.0
Very true of me now	42.0
I am concerned about how my facilitation affects the attitudes of those directly involved in the use of "Inclusion"	
Irrelevant	6.5
Not true of me now	8.1
Somewhat true of me now	25.8
True of me now	17.7
Very true of me now	41.9
I would like to know more about "Inclusion"	
Irrelevant	0.0
Not true of me now	6.4
Somewhat true of me now	30.7
True of me now	11.3
Very true of me now	51.7
I need more information about and understanding of "Inclusion"	
Irrelevant	1.6
Not true of me now	9.6
Somewhat true of me now	29.0
True of me now	12.9
Very true of me now	46.7
I would like to determine how to enhance my facilitation skills	
Irrelevant	1.6
Not true of me now	8.1
Somewhat true of me now	29.1
True of me now	21.0
Very true of me now	40.4
I am concerned about being held responsible for facilitating use of "Inclusion"	
Irrelevant	1.6
Not true of me now	16.2
Somewhat true of me now	22.6
True of me now	24.2
Very true of me now	35.4
Currently, other priorities prevent me from focusing my attention on "Inclusion"	
Irrelevant	6.5
Not true of me now	16.2
Somewhat true of me now	32.2
True of me now	19.4
Very true of me now	25.8
I would like to know where I can learn more about "Inclusion"	
Irrelevant	3.2
Not true of me now	16.1
Somewhat true of me now	27.5
True of me now	14.5
Very true of me now	38.8

Collaboration Leadership Survey Items	Percent
I would like to develop working relationships with administrators and other change facilitators to facilitate the use of "Inclusion"	
Irrelevant	1.6
Not true of me now	9.7
Somewhat true of me now	30.6
True of me now	19.4
Very true of me now	38.7
I am concerned about criticism of my work with "Inclusion"	
Irrelevant	19.4
Not true of me now	29.0
Somewhat true of me now	27.4
True of me now	12.9
Very true of me now	11.3
Working with administrators and other change facilitators in facilitating use of "Inclusion" is important to me	
Irrelevant	1.2
Not true of me now	12.9
Somewhat true of me now	24.2
True of me now	16.1
Very true of me now	45.2
I wonder whether use of "Inclusion" will help or hurt my relations with my colleagues	
Irrelevant	3.2
Not true of me now	16.1
Somewhat true of me now	27.5
True of me now	14.5
Very true of me now	38.7

(table continued)

Collaboration Leadership Survey Items	Percent
I would like to coordinate my efforts with other change facilitators	
Irrelevant	1.6
Not true of me now	4.8
Somewhat true of me now	40.3
True of me now	16.1
Very true of me now	37.1
I want to know what priority my superiors want me to give this innovation	
Irrelevant	0.0
Not true of me now	12.9
Somewhat true of me now	21.0
True of me now	19.4
Very true of me now	46.8
I would like to help others in facilitating the use of "Inclusion"	
Irrelevant	6.5
Not true of me now	37.1
Somewhat true of me now	19.4
True of me now	11.3
Very true of me now	25.8
I see a potential conflict between facilitating "Inclusion" and overloading staff	
Irrelevant	4.8
Not true of me now	9.7
Somewhat true of me now	21.0
True of me now	16.1
Very true of me now	48.4
I am concerned about how my facilitating the use of "Inclusion" affects those directly involved in the use of it	
Irrelevant	1.6
Not true of me now	14.6
Somewhat true of me now	35.5
True of me now	24.2
Very true of me now	24.2
Communication and problem solving relative to "Inclusion" affects those directly involved in the use of it	
Irrelevant	8.1
Not true of me now	32.3
Somewhat true of me now	33.9
True of me now	17.7
Very true of me now	8.0
I would like to modify my mode of facilitating the use of "Inclusion" based on the experiences of those directly involved in its use.	
Irrelevant	6.5
Not true of me now	14.5
Somewhat true of me now	37.1
True of me now	19.4
Very true of me now	22.6
I would like to familiarize other departments or persons with the progress and process of facilitating the use of "Inclusion"	
Irrelevant	11.3
Not true of me now	24.2
Somewhat true of me now	27.4
True of me now	19.4
Very true of me now	17.8

Supports for Staff and Students Items	Percent
I am concerned because responding to the demands of staff relative to "Inclusion" takes so much time	
Irrelevant	3.2
Not true of me now	11.3
Somewhat true of me now	32.2
True of me now	11.3
Very true of me now	42.0
I am preoccupied with things other than "Inclusion"	
Irrelevant	3.2
Not true of me now	12.9
Somewhat true of me now	32.3
True of me now	11.3
Very true of me now	40.3
I am concerned about facilitating use of "Inclusion" in view of limited resources	
Irrelevant	3.2
Not true of me now	8.1
Somewhat true of me now	32.2
True of me now	21.0
Very true of me now	35.5
I would like to know what resources are necessary to adopt "Inclusion"	
Irrelevant	0.0
Not true of me now	11.3
Somewhat true of me now	29.1
True of me now	21.0
Very true of me now	38.7
I am concerned about finding and allocating time needed for "Inclusion"	
Irrelevant	1.6
Not true of me now	22.5
Somewhat true of me now	30.7
True of me now	14.5
Very true of me now	30.6

Results

Research on needs and concerns of administrators as change facilitators when preparing, implementing and maintaining successful inclusion programs was the focus of the research. The National Survey on Inclusive Education (1994) found seven factors necessary for successful inclusion programs: (a) visionary leadership, (b) collaboration, (c) supports for staff and students, (d) refocused use of assessment, (e) funding, (f) effective parental involvement, and (g) models and classrooms practices that support inclusion (NCERI, 1994).

Three of the seven factors were tested for significance by the assistant principals in this study. The authors postulated these three factors of (1) visionary leadership, (2) collaboration, and (3) supports for staff and students held greatest significance for the assistant principals. These survey factors are related directly to

BEST COPY AVAILABLE

school administrator roles and responsibilities for placement decisions of students with disabilities.

Paired sample *t*-tests were used to determine if there were significant differences between the means of the three subsections. The alpha level for these analyses was set at .017 using the Bonferroni correction for multiple comparisons. The means on the supports for staff and students and visionary leadership were not statistically significant (t = .999, p = .322). The means on collaboration and supports for staff (t = 3.998, $p < .0001$, r = .45) and students and collaboration and visionary leadership (t = 5.664, $p < .0001$, r = .58) were statistically significant. Effect sizes were calculated to aid interpretation of results for practical significance of the research. Medium effect sizes were determined for collaboration and supports and for collaboration and visionary leadership. The mean scores and standard deviations for each subsection are shown in the Table 2.

Table 2
Means and Standard Deviations of Assistant Principals' Perceptions of Visionary Leadership, Collaboration, and Supports for Staff and Students Subsections

Subsection	*n*	*M*	*SD*
Visionary Leadership	62	4.74	1.86
Collaboration	62	4.00	1.93
Supports for Staff and Students	62	4.59	1.86

Note. Maximum score = 7. Minimum score = 0.

Table 3 shows the compilation of frequency counts in relation to the three subsections taken from the survey results. The results showed that 58% of the sampled assistant principals indicated that concerns in the area of visionary leadership were true or very true of their present perceptions. Fifty-two percent of the sample indicated that their concerns in the area of collaboration were true or very true of their present perceptions, and 53% felt that concerns in the area of supports for staff and students were true or very true of them. These results indicated that approximately one-third, or 36%, of the participants felt that each subsection was "very true" of their present concerns, while more than 18% of the participants felt that each subsection was "true" of their concerns. In summary, over one-half, or 54%, indicated their present level of concern for each of the three factors was true or very true.

The recognition of visionary leadership, the first factor, as essential for successful inclusion to take place demonstrated an awareness by the assistant principals that inclusion must be viewed as part of a larger picture that is important to the overall mission of the school. Villa and Thousand (1995) defined visionary leaders as leaders in inclusive education who clarify a vision of school success that suggests "(1) all children are able to learn, (2) all children should be educated together in their community schools, and (3) the school system is responsible for addressing the unique needs of all children" (p. 59).

Table 3
Compilation of Frequency Counts of Assistant Principals' Responses to Subsections of Concerns Questionnaire for Change Facilitators

Subsection	Question Response	Percent
Visionary Leadership		
	Irrelevant	3
	Not true of me now	11
	Somewhat true of me now	29
	True of me now	18
	Very true of me now	40
Collaboration		
	Irrelevant	6
	Not true of me now	21
	Somewhat true of me now	21
	True of me now	20
	Very true of me now	32
Supports for Staff & Students		
	Irrelevant	2
	Not true of me now	13
	Somewhat true of me now	31
	True of me now	16
	Very true of me now	37

The second factor cited by surveyed assistant principals was a need for collaboration. Rosenholtz (1989) described collaborative schools as schools where the educational processes are viewed as collective rather than individual enterprises and where norms and opportunities for continuous improvement and career-long learning not only exist but are supported and nurtured by the school culture. Such a definition has great merit in the inclusive schools. For too long, students with special needs have been viewed as the responsibility of the special education teacher, and any involvement of the students with the general classroom teachers and students was seen as a privilege rather than a right. True collaboration gives all members of the school family a sense of efficacy as they are encouraged to participate in the totality of the school setting.

The third factor delineated by the assistant principals in this study was the need for supports for staff and students. This was a strong indicator of the assistant principals' awareness of the need for more appropriate training and follow-up in the implementation of this paradigm shift at the school level. Not only were they asking for help and support in their role in this process, but their concern extended to the needs of students and staff to create the inclusive school that meets the needs of all children of the school community.

The role of the assistant principal as it presently exists is varied; however, current literature frequently links the assistant principal to the role of the school disciplinarian (Marshall, 1993). A recent study by Pool and Petrie (1996) suggested that assistant principals viewed their roles as much more important to the success of the school than that of disciplinarian. Their research indicated that while the assistant principals recognized and accepted discipline as a major part of their role, nonetheless, they believed they were being underutilized in terms of the potential service they could provide to the overall program of the school through greater involvement with teachers and students to create stronger, more successful educational experiences within their schools. Our research suggested that this perception was common among the assistant principals in our study as well.

The assistant principals in the Academy suggested a concern about their readiness to collaborate with others in the implementation of the inclusion process. Their level of concern suggested they did not feel knowledgeable enough about the implementation process to be secure in their ability to make appropriate contributions to the success of the program. The data did not indicate that they do not support inclusion as an ideal; rather, the data suggested they had an awareness of their need for greater knowledge, support and development in how to help bring about successful implementation.

Discussion

Although inclusion is not a direct federal mandate, inclusive special education delivery is gaining support and popularity. Wayne Qualls, Commissioner of the Tennessee State Department of Education, requested reviews and reactions of "Excellence in Education Through Inclusion: A Vision for the Twenty-first Century" (Personal communication, April 22, 1994). This draft served as a guide for State Department staff to develop strategies for the provision and enhancement of educational opportunities through inclusion. The collaborative efforts of Tennessee Department of Education, Division of Special Education, Divisions of Curriculum and Instruction and Vocational Education created the draft to develop "a plan designed to move toward an inclusionary system of education" (personal communication, April 22, 1994).

Successful transformation of inclusion of students with special needs in the general classroom requires proper training and support for all who are involved. Lieberman and Miller (1986) stated that "mandating new policy without attending to organizing, supporting, and providing teachers and principals with the necessary learnings they need to carry out any school improvement efforts will be ineffective" (p. 100). Implementing successful inclusion programs and implementing successful school reform are closely linked.

Supporters of school reform and supporters of inclusion agree that all students should be educated as full members of the school (NASBE, 1994). This practice requires a collaborative effort among all instructional staff members to determine the most appropriate education to meet the individual needs of all learners. Another overlapping practice of school reform and inclusion is developing school autonomy. Building-based change encourages more input from key players, or stakeholders, to make changes and decisions based on the culture of the school (NASBE, 1994). In support, Cunningham and Gresso found "when employees have an opportunity to be self-directed in their learning, they are likely to be highly motivated and committed to their development" (1993, p. 189).

The Concerns Questionnaire attempted to find areas of concern and interest for assistant principals linking selected factors to successful inclusion programs. Participants indicated strong concerns in visionary leadership, collaboration, and supports for staff and students.

Katsiyannis, Conderman, and Franks (1996) recommended and cautioned that:

> 1. Inclusionary practices should be carefully planned to avoid forcing inclusion on all students regardless of individual needs. Such programming should occur only after the necessary supports are in place.
> 2. Financial support, inservice training, and technical assistance are necessary components for effective inclusionary practices.
> 3. Barriers to inclusion such as special education reimbursement formulas, limited training opportunities, categorical teacher training and certification requirements must be addressed.

4. Passions for or against inclusion may interfere with the provision of an appropriate education for students with disabilities, and therefore they should be monitored.
5. Inclusionary practices must result in documented benefits for students with disabilities and their same age peers.
6. Program evaluation and ongoing/empirical research are necessary components of inclusionary programming. (pp. 84-85)

Critics of special education note that pullout delivery programs are expensive, encourage segregation, and fragment education of students with special needs. Madeline Will (1986) introduced the regular education initiative to promote a merger of special education and regular education governance and funding. Inclusion addressed instructional delivery, methods, and strategies by having the support services provided in the general education classroom. Relocating services to the general education class is a paradigm switch from the traditional way of educating students with disabilities. Any successful change requires appropriate, supportive training and development. Because assistant principals are frequently assigned the responsibility of placement of students in special education, there is a need for specialized and ongoing development of their abilities to implement these changes to best meet the needs of students and teachers. Professional development opportunities that focus on their identified concerns in the areas of visionary leadership, collaboration, and supports for staff and students would be welcomed by the assistant principals, and consequently, should provide more opportunities for the development of successful inclusive schools.

References

Cunningham, W. G., & Gresso, D. W. (1993). *Cultural leadership: The culture of excellence in education.* Needham Heights, MA: Allyn & Bacon.

Fuchs, D., & Fuchs, L. S. (1995). What's 'special' about special education? *Phi Delta Kappan, 76*(7), 522-529.

Fullan, M. (1996). *Change forces probing the depths of educational reform.* Cambridge, UK: Falmer Press.

Gay, L. R. (1996). *Educational research: Competencies for analysis and application.* Englewood Cliffs, NJ: Merrill.

Goodlad, J. I. (1990). *Places where teachers are taught.* San Francisco, CA; Jossey-Bass.

Hall, G. E., Newlove, B. W., George, A. A., Rutherford, W. L., & Hord, S. M. (1991). *Measuring change facilitator stages of concern: A manual for the use of the CFSoCQ questionnaire.* Greeley, CO: Center for Research on Teaching and Learning, University of Northern Colorado.

Hittleman, D. R., & Simon, A. J. (1992). *Interpreting educational research.* New York: Macmillan Publishing Company.

Katsiyannis, A., Conderman, G., & Franks, D. (1996). Students with disabilities: Inclusionary programming and the school principal. *NASSP Bulletin, 80*(578), 81-85.

Keaster, R. (1996, November). *Administrators' attitudes and the change process: Implications for inclusive education.* Paper presented at the annual meeting of Southern Regional Council for Educational Administration, Savannah, GA.

Lieberman, A., & Miller, L. (1986). School improvement: Themes and variations. In A. Lieberman (Ed.), *Rethinking School Improvement* (pp. 96-111). New York, NY: Teachers College Press.

Lombardi, T. (1994). *Responsible inclusion of students with disabilities.* Bloomington, IN: Phi Delta Kappa.

MacKay, L. (1994, November). *The dynamics of succeeding with inclusion.* Paper presented at the meeting of the Assistant Principal's Academy, Abingdon, VA.

Marshall, C. (1993). *The unsung role of career assistant principals.* In Reports/Research/Technical (143). (ERIC Document Reproduction Services, No. ED 355 653)

Murphy, J., & Schiller, J. (1996). *Transforming America's schools as administrators' call to action.* Chicago: Open Court.

NASBE says teachers ready for inclusion. (1994, January). *Inclusive Education Programs, 1*(1), 12.

National Center on Educational Restructuring and Inclusion (NCERI). (1994). *National survey on inclusive education.* New York: The City University of New York, The Graduate School and University Center.

Pool, H., & Petrie G. (1996, January). *The assistant principalship in Georgia: Can we narrow the gap between ideal and practice?* Paper presented at the Southern Regional Council on Educational Administration Annual Conference, Savannah, Georgia.

Pues, S. (1990). Adults with special learning needs: An overview. *Adult Learning, 2*(2), 1720.

Rogers, J. (1993, May). The inclusion revolution. *Research Bulletin, 11*, 14.

Rosenholtz, S. (1989). *Teachers' workplace: The social organization of schools.* New York: Longman.

Stainback, W. M., & Stainback, S. (1984). A rationale for the merger of special and regular education. *Exceptional Children, 51*(2), 102-111.

Stainback, W. M., & Stainback, S. (1985). *Integration of students with severe handicaps into regular schools.* Reston, VA: The Council for Exceptional Children.

Villa, R. A., & Thousand, J. S. (1995). *Creating an inclusive school.* Alexandria, VA: Association for Supervision and Curriculum Development.

Will, M. C. (1986). Educating students with learning problems: A shared responsibility. *Exceptional Children, 52* (5), 411-416.

Appendix

Concerns Questionnaire for Change Facilitators

The purpose of this questionnaire is to determine what you are thinking about regarding your responsibilities as a change facilitator for an innovation. It is not necessarily assumed that you have change facilitator responsibilities. This questionnaire is designed for persons who do not serve as change facilitators as well as for those who have major responsibility for facilitating change. Because the questionnaire attempts to include statements that are appropriate for widely diverse roles, there will be items that appear to be of little relevance or irrelevant to you at this time. For the completely irrelevant items, please circle "0" on the scale. Other items will represent those concerns you do have, in varying degrees of intensity, and should be marked higher on the scale.

For example:

This statement is very true of me at this time.	0 1 2 3 4 5 6 7
This statement is somewhat true of me now.	0 1 2 3 4 5 6 7
This statement is not true of me at this time.	0 1 2 3 4 5 6 7
This statement seems irrelevant to me.	0 1 2 3 4 5 6 7

Please respond to the items in terms of your present concerns, or how you feel about your involvement with facilitating "*Inclusion*." We do not hold to any one definition of this program, so please think of it in terms of your own perceptions of what it involves. Remember to respond to each item of your present concerns about your involvement or potential involvement as a facilitator of "*Inclusion*."

Thank you for taking time to complete this task. Please feel free to write any comments, reactions or questions you may have about the items on the questionnaire. Also, use the last page to express any additional concerns you have about "*Inclusion*" or this questionnaire.

Note(s): From: G. E. Hall, B. W. Newlove, A. A. George, W. L. Rutherford, & S. M. Hord. (1991). *Measuring change facilitator stages of concern: A manual for the use of the CFSoCQ questionnaire.* Copyright 1989 by Concerns Based Systems International.

0	1	2	3	4	5	6	7
Irrelevant	Not true of me now		Somewhat true of me now			Very true of me now	

1. I would like more information about the purpose of "Inclusion"	0 1 2 3 4 5 6 7
2. I am more concerned about facilitating use of "Inclusion"	0 1 2 3 4 5 6 7
3. I would like to develop working relationships with administrators and other change facilitators to facilitate the use of "Inclusion"	0 1 2 3 4 5 6 7
4. I am concerned because responding to the demands of staff relative to "Inclusion" takes so much time	0 1 2 3 4 5 6 7
5. I am not concerned about "Inclusion" at this time	0 1 2 3 4 5 6 7
6. I am concerned about how the facilitation affects the attributes for those directly involved in the use of "Inclusion"	0 1 2 3 4 5 6 7
7. I would like to know more about "Inclusion"	0 1 2 3 4 5 6 7
8. I am concerned about criticism of my work with "Inclusion"	0 1 2 3 4 5 6 7
9. Working with administrators and other change facilitators in facilitating use of "Inclusion" is important to me	0 1 2 3 4 5 6 7
10. I am preoccupied with things other than "Inclusion"	0 1 2 3 4 5 6 7
11. I wonder whether use of "Inclusion" will help or hurt my relations with my colleagues	0 1 2 3 4 5 6 7
12. I need more information about and understanding of "Inclusion"	0 1 2 3 4 5 6 7
13. I am thinking that "Inclusion" could be modified in view of limited resources	0 1 2 3 4 5 6 7
14. I am concerned about facilitating use of "Inclusion" in view of limited resources	0 1 2 3 4 5 6 7
15. I would like to coordinate my efforts with other change facilitators	0 1 2 3 4 5 6 7
16. I would like to know what resources are necessary to adopt "Inclusion"	0 1 2 3 4 5 6 7

BEST COPY AVAILABLE

17. I want to know what priority my superiors want to give this innovation 0 1 2 3 4 5 6 7
18. I would like to excite those directly involved in the use of "Inclusion" about their part in it 0 1 2 3 4 5 6 7
19. I am considering use of another innovation that would be better than the one that is currently being used 0 1 2 3 4 5 6 7
20. I would like to help others in facilitating the use of "Inclusion" 0 1 2 3 4 5 6 7
21. I would like to determine how to enhance my facilitation skills 0 1 2 3 4 5 6 7
22. I spend little time thinking about "Inclusion" 0 1 2 3 4 5 6 7
23. I see a potential conflict between facilitating "Inclusion" and overloading staff 0 1 2 3 4 5 6 7
24. I am concerned about being held responsible for facilitating use of "Inclusion" 0 1 2 3 4 5 6 7
25. Currently, other priorities prevent me from focusing my attention on "Inclusion" 0 1 2 3 4 5 6 7
26. I know of another innovation that I would like to see used in place of "Inclusion" 0 1 2 3 4 5 6 7
27. I am concerned about how my facilitating the use of "Inclusion" affects those directly involved in the use of it 0 1 2 3 4 5 6 7
28. Communication and problem-solving relative to "Inclusion" take too much time 0 1 2 3 4 5 6 7
29. I wonder who will get the credit for implementing "Inclusion" 0 1 2 3 4 5 6 7
30. I would like to know where I can learn more about "Inclusion" 0 1 2 3 4 5 6 7
31. I would like to modify my method of facilitating the use of "Inclusion" based on the experiences of those directly involved in its use 0 1 2 3 4 5 6 7
32. I have alternative innovations in mind that I think would better serve the needs of our situation 0 1 2 3 4 5 6 7
33. I would like to familiarize other departments or persons with the progress and process of facilitating the use of "Inclusion" 0 1 2 3 4 5 6 7
34. I am concerned about finding and allocating time needed for "Inclusion" 0 1 2 3 4 5 6 7
35. I have information about another innovation that I think would produce better results that the one we are presently using 0 1 2 3 4 5 6 7

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 1, 67-72

Proper Use of the Two-Period Crossover Design When Practice Effects are Present

M. Suzanne Moody
Auburn University

In a two-period crossover design, counterbalancing does not remove carryover effects but rather completely entangles them with treatment effects. Not only does this entanglement result in a loss of statistical power, but more importantly, it jeopardizes proper interpretation of results. The major controversy surrounding the use of this design is discussed, and a case is made for the proper use of this design in educational research. Labeled plots, a discussion of terminology, and SPSS commands are provided to aid in identification of the three main effects (treatment, session, and order) and in interpretation of results. Keppel's procedure for removal of variation due to practice effects and test for treatment effect is shown by example to be equivalent to a test for all three effects, including differential carryover (i.e., order effect).

The two-period crossover design, also termed the replicated 2 X 2 Latin square design, is one in which "one group of subjects receives Treatment A for one period of time and Treatment B for the next period of time, while a second group receives Treatment B for the first period and Treatment A for the second period" (Cotton, 1989, p. 503). This counterbalanced, repeated measures design is sometimes used in educational research with the intention of rotating out any differences between groups when intact groups must be used. For example, "if one group should be more intelligent on the average than the other, each treatment would benefit from this superior intelligence" (Ary, Jacobs, & Razavieh, 1990, p. 341). The crossover design is also used because, as with other repeated measures designs, each subject acts as his or her own control, thereby removing subject effects from the error variance and increasing statistical power (Cotton, 1989; Grizzle, 1965). In addition, some authors of popular texts recommend the use of this design in cases in which practice effects are present and cannot be eliminated (Borg & Gall, 1989; Keppel, 1991; Winer, Brown, & Michels, 1991). However, other authors insist that a crossover design should not be used unless carryover effects such as practice are not present or can be eliminated (Ary et al., 1990; Lentner & Bishop, 1993). It will be the purpose of this discussion to further examine the apparent controversy concerning the proper use of crossover designs when practice effects are present and to offer suggestions for analyzing the data if one chooses to employ this design.

The Problem

Ary et al. (1990) state that the counterbalanced design "should be used only when the experimental treatments are such that exposure to one treatment will have no effect on subsequent treatments," and admits that "this requirement may be hard to satisfy in much of educational research" (p. 341). To the contrary, Keppel (1991) believes that if "the equal occurrence of each experimental treatment at each stage of practice" is ensured (p. 335), such a design is a solution to the problem of practice and treatment effect entanglement. According to Keppel, the problem with practice effects occurs only when there is the possibility of *differential* carryover effects. Differential carryover effects (also known as asymmetrical transfer effects) occur when the amount of carryover due to practice for the second period of testing is different for Treatment A than it is for Treatment B. Differential carryover effects are equivalent to a testing session by treatment interaction. A testing session by treatment interaction is one of three interactions generated by the two-period crossover design. The other two interactions are a treatment by order interaction that is equivalent to a session main effect and a session by order interaction that is equivalent to a treatment main effect. Obviously, session, order, and treatment effects are each entangled with each other. As a result, if the test for differential carryover effects (i.e., testing session by treatment interaction) is significant, then the test for either of the other two main effects is not interpretable, because there is no way to determine how

---

M. Suzanne Moody is a Ph.D. candidate in Educational Psychology, Measurement and Statistics, Department of Educational Foundations, Leadership, and Technology at Auburn University. Please direct all correspondence to the author at EFLT, Haley Center 4036, Auburn University, AL 36849 or e-mail: moodysu@mail.auburn.edu.

much of the increase in scores for the second testing is attributable to practice and how much is attributable to treatment.

Terminology

As Cotton (1989) has noted, several authors have mistakenly equated differential carryover effects with a treatment by order interaction (i.e., session main effect). He suggests that this confusion may have been the result of inconsistent language use. The present author believes that ambiguous language may also give the aforementioned advice concerning the use of crossover designs an appearance that it is more contradictory than it really is. To avoid further confusion, it will be necessary to list all of the alternative terms used in the literature to represent the words *session*, *treatment*, *order*, and *differential carryover*. Session is also sometimes termed period, position, trial, or stage. Treatment is also sometimes termed condition. Order is also sometimes termed sequence. Differential carryover is sometimes termed asymmetric transfer or residual effects, but more confusingly, it is sometimes shortened to carryover. However, the term carryover is usually reserved for any influential transfer (whether differential or symmetrical) that occurs from the first testing session to the second and can be due to such things as practice, fatigue, or a drug that has not completely cleared the body's system. Consider how these terms are used in the following example. In a study by the present author, the Mental Rotations Test (Vandenberg & Kuse, 1978) was administered twice to 34 females with the purpose of testing whether females perform significantly different during the menstrual phase of their menstrual cycle than during the luteal phase. During the first testing, half of the females (Group A) were in the luteal phase of their menstrual cycle and the other half (Group B) were in their menstrual phase. During the second testing, the 17 females of Group A were in their menstrual phase and the 17 in Group B were in their luteal phase. Each testing is a session. The luteal and menstrual phases are the treatments. Whether the woman received the menstrual treatment or the luteal treatment during the first testing session is the order. If females who are in their luteal phase during the first testing session have greater increases in their scores for the second testing session than do women who are in their menstrual phase during the first session or vice versa, then the carryover effect (i.e., practice effect) is differential. In more general terms, differential carryover would be present if, for some reason, "undergoing" the first testing session caused the groups to be systematically different as they go into the second testing session.

Tests for Main Effects

Keppel (1991) recommends the examination of a plot for possible interaction between testing position (i.e., session or period) and treatments and an abandonment of within subject analyses unless this interaction is absent. Hills and Armitage (1979), Cotton (1989), and Jones and Kenward (1989) clearly demonstrate an actual F test for a treatments by session interaction (i.e., differential carryover) and provide the mathematical underpinnings. Unlike Keppel, these authors do not require the stipulation of equal n's for the Groups A and B. Additionally, Ratkowsky, Evans, and Alldredge (1993) list the commands used in the statistical computing package SAS for the procedure. The following commands are the ones that may be used to carry out the procedure in SPSS, a package popularly used in educational research.

```
MANOVA session1 session2  BY  order(0 1)
/WSFACTORS session(2)
/METHOD UNIQUE
/ERROR WITHIN+RESIDUAL
/PRINT SIGNIF(MULT AVERF EFSIZE)
/NOPRINT PARAM(ESTIM).
```

Notice that the data are actually analyzed in a split plot design with session as the within subjects factor and testing order as the between subjects factor. The data input coding and output are presented in Tables 1 and 2 respectively. The plots of the three associated interactions (i.e., main effects) are presented in Figures 1, 2, and 3.

Interpretation of Results

The results indicate a significant practice effect (session effect) with $p < .0001$ but a nonsignificant order effect (i.e., differential practice carryover effect) with $p = .74$. However, because the test for differential carryover is between subjects as opposed to within subjects, it is necessary to consider the lack of significance merely the result of a lack of statistical power due to a small sample size (Jones & Kenward, 1989). Herein lies the most probable cause of the conflicting advice mentioned in the introduction of this discussion. Opponents of the use of two-period crossover designs when practice effects are present point to the lack of power in making the determination of significant differential carryover effects, and the misinterpretations that result, as one of their major objections to the design. Brown (1980) has shown that, even setting alpha at .10 as suggested by Grizzle

(1965), the number of subjects required to obtain sufficient power for the between subjects test for differential carryover negates the main advantage of the crossover design. In addition, if the test for differential carryover is determined to be significant, then the researcher is forced to resort to a between subjects test for main treatment effect using the first session data only. Here again, the small sample size often employed in the crossover design will likely generate insufficient power to detect a significant between subjects difference in treatments and render the study a waste of time and resources.

Table 1
Data Input Coding

SESSION1	SESSION2	ORDER
4.00	2.00	.00
7.00	22.00	.00
17.00	26.00	.00
10.00	21.00	.00
12.00	14.00	.00
5.00	19.00	.00
4.00	22.00	.00
11.00	18.00	.00
8.00	12.00	.00
6.00	18.00	.00
22.00	28.00	.00
4.00	11.00	.00
11.00	13.00	.00
10.00	26.00	.00
10.00	24.00	.00
9.00	14.00	.00
14.00	20.00	.00
11.00	14.00	1.00
8.00	10.00	1.00
12.00	12.00	1.00
13.00	29.00	1.00
11.00	16.00	1.00
15.00	13.00	1.00
16.00	10.00	1.00
12.00	22.00	1.00
22.00	28.00	1.00
13.00	14.00	1.00
9.00	2.00	1.00
11.00	10.00	1.00
25.00	33.00	1.00
8.00	4.00	1.00
19.00	19.00	1.00
12.00	16.00	1.00
8.00	19.00	1.00

BEST COPY AVAILABLE

Table 2
SPSS Output

```
******Analysis of Variance******

          34 cases accepted.
           0 cases rejected because of out-of-range factor values.
           0 cases rejected because of missing data.
           2 non-empty cells.

           1 design will be processed.

Tests of Between-Subjects Effects.

 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation          SS      DF        MS       F    Sig of F

 WITHIN+RESIDUAL         2036.12      32     63.63
 ORDER                      7.12       1      7.12     .11       .740
-----------------------------------------------------------------------
 Effect Size Measures and Observed Power at the .0500 Level
                         Partial      Noncen-
 Source of Variation     ETA Sqd      trality     Power

 ORDER                    .003         .112       .053
-----------------------------------------------------------------------
Tests involving 'SESSION' Within-Subject Effect.

 Tests of Significance for T2 using UNIQUE sums of squares
 Source of Variation          SS    DF      MS         F      Sig of F

 WITHIN+RESIDUAL         562.82    32    17.59
 SESSION                 542.12     1   542.12     30.82        .000
 ORDER BY SESSION        147.06     1   147.06      8.36        .007
-----------------------------------------------------------------------
 Effect Size Measures and Observed Power at the .0500 Level
                          Partial     Noncen-
 Source of Variation      ETA Sqd     trality    Power
 SESSION                    .491       30.823    1.000
 ORDER BY SESSION           .207        8.361     .799
```

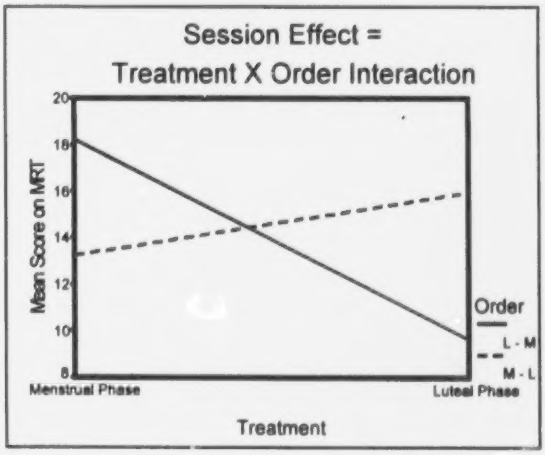


Figure 1.

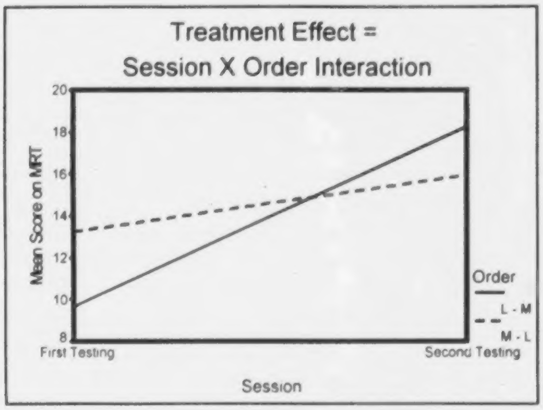


Figure 2.

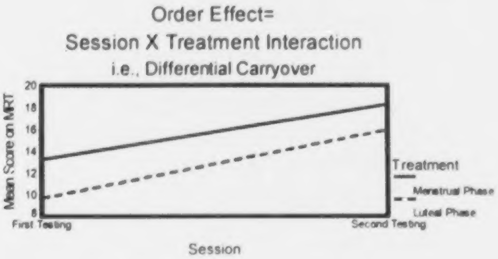


Figure 3.

These potential problems might cause researchers rightfully to shy away from using crossover designs. However, if an educational researcher does not anticipate differential carryover and elects to employ the design, he/she should nonetheless be sure to document evidence of the lack of such effects. If the level of.power for the test of differential practice carryover leaves one in doubt, other factors should be considered, such as the parallelism (or lack thereof) of the treatment by session plot as recommended by Keppel (1991) and the similarity of the magnitudes of the mean practice effect for the entire group of subjects to the magnitudes found in the literature. In this example, the plot is fairly parallel (i.e., ordinal) and the mean increase of 5.65 points due to practice for all the subjects is similar to the one of 5.78 points obtained from females by Peters, Laeng, Latham, Jackson, Zaiyouna, and Richardson (1995) with no treatment applied. T. Holmes (personal communication, November 17, 1996) suggests that an additional piece of evidence of negligible differential carryover is a finding of "little difference between the treatment main effect (which is averaged across both sessions) and the specific effect between first-session treatment means . . . [which] cannot contain carryover effects." The data of the present example yield means that differ by only 0.647 points. Therefore, because one can be fairly confident in denying the significance of the treatment by session interaction in this example, the phase (treatment) effect (i.e., order by session interaction) may now and only now be meaningfully interpreted as significant ($p < .05$). Moreover, with the finding of a significant treatment effect, the researcher may engage in a discussion of the importance of the finding as evidenced by the effect size.

Less Trouble Than Keppel Thought

Keppel (1982) presents a method for removing variation due to practice in order to make crossover designs more statistically sensitive (i.e., powerful). However, Keppel (1991) warns of "the need to [graphically] check on the possibility of an interaction between treatments and testing position" prior to performing his procedure and that "any suspicious departure from parallel functions---that is, any interaction---should be assessed statistically by means of a fairly complicated statistical analysis which [he] will not consider" (p. 364). Interestingly, in the case of the 2 X 2 Latin square design, Keppel's procedure for removing variation due to practice and the test for the treatment by session interaction which Keppel said he would not consider in his article are actually one in the same. A standard within subjects analysis applied to the data of the present example with phase as the within subjects factor yields $p = .044$ and an effect size of .117. Applying Keppel's procedure to the data, we first obtain a total mean practice effect of 5.647. We next obtain a mean practice effect for Group A of 8.588 and 2.7059 for Group B. Therefore, the correction constant for Session 1 is 8.588 - 5.647 = 2.941 and 2.7059 - 5.647 = -2.941 for Session 2. An adjustment is made to the original scores by adding 2.941 to all the scores from Session 1 and -2.941 to all the scores from Session 2. A standard within subjects analysis using the adjusted scores with phase as the within subjects factor and the *df* for the error term reduced by one yields $F = 8.36$ ($p = .007$) and an effect size of .207 for the treatment (phase) effect. As intended, the adjustments result in a much more powerful test. Note, however, that if the unadjusted scores are analyzed as shown previously in a split plot design with session instead of phase (i.e., treatment) as the within subjects factor and testing order as a between subjects factor, the values obtained are exactly the values obtained using Keppel's procedure but with the added benefit of an F test for differential carryover (i.e., order effect).

More Trouble Than Some Others Thought

The discussion of and test for differential carryover is lacking in educational research literature. For example, in their text, Borg and Gall (1989, p. 709) offer as an example of an appropriate use of a counterbalanced experiment one in which, for the purpose of investigating the effects of text difficulty on reading rate, half the subjects were randomly assigned to read an eighth-grade passage first and an eleventh-grade passage second. The other half read the eleventh-grade passage first and the eighth-grade passage second. Though one would suspect the possibility of differential carryover in this experiment, Borg and Gall do not mention this possibility. By claiming that "counterbalanced designs are used to avoid the problems of interpretation due to [what they mislabel as] order effects" (p. 708), the authors of this popular text may be leading some to believe that counterbalancing is sufficient for avoiding the problems caused by practice effects in a repeated measures design. As may be gathered from the previous discussion, such thinking is fallible. The consequences of failing to understand the fundamentals of this design are exemplified in the following two studies.

In a study of learning strategy training, Dansereau, Brooks, Holley, and Collins (1983) trained one group of participants in text-oriented strategies during the first half of a semester and in management of concentration strategies during the second half. Another group received the opposite training sequence. Text-processing tasks served as dependent measures during each half semester. It was assumed that the strategies learned during the first half of the semester were carried over and used along with the strategies introduced in the second half. The researchers report that participants who received the text-oriented strategies prior to the concentration strategies benefited more than those who received the strategies in reverse order. The researchers explain that one possibility for the results is that text-oriented strategies may require more time to master and that participants who received these strategies first had the entire semester to practice them while the other group had only the second half of the semester to master these techniques. However, possible explanations for the results and the need for further research suggested by the researchers are unnecessary, for the researchers mistakenly derived the significant training order effect from a simple between subjects comparison of dependent measure scores in the second half of the semester instead of from a test for testing session by treatment interaction. It appears that the researchers made a mistake similar to the one reported by Cotton (1989) which is the mistake of equating differential carryover effects with a treatment by order interaction. Although the present author does not have access to the raw data, a session by treatment plot of means appears almost parallel, thereby giving no indication of an order (i.e., differential carryover) effect. Furthermore, given a lack of order effects and what appears to be a testing session by order interaction, the researchers could have interpreted the concentration strategies as being more effective than the text-oriented strategies had they been interested in testing such a hypothesis. However, they were incorrect in reporting a significant training order effect.

Sensenig, Mazeika, and Topf (1989) conducted a study in which participants in one group were taught to read words using flash cards during a first session and taught to read another set of words (equivalent to the first in difficulty) using both flashcards and sign language during a second session. For another group, the order was reversed for the two sessions. The researchers report that "learning to read with accompanying sign significantly increased reading performance" (p. 124). Unfortunately, the researchers failed to consider the possibility of differential carryover. Perhaps, without the researchers' knowledge, the group that had been taught the sign language strategy during the first session could have mentally employed this strategy on the second set of words during the second session. Regardless of the reason, the nonparallel result of a session by treatment plot of means indicates that such an order effect is likely. Therefore, due to the entanglement of "practice" and treatment effects, the treatment differences from the repeated measures analysis should not have been interpreted as statistically significant.

In conclusion, the present author would like to re-emphasize the risks associated with the crossover design. The design is not being either recommended or condemned here, but rather proper methods for analysis are being suggested to researchers who choose to employ the design notwithstanding these risks. In addition, it is suggested that, even at an introductory level, instructors of educational research courses and authors of educational research texts need to take due care in presenting the topic of counterbalanced repeated measures designs so as to prevent naive implementation of these designs in educational research.

References

Ary, D., Jacobs, L. C., & Razavieh, A. (1990). *Introduction to research in education.* Fort Worth, TX: Holt, Rinehart and Winston.

Borg, W. R., & Gall, M. D. (1989). *Educational research: An introduction.* New York: Longman.

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics, 36*, 69-79.

Cotton, J. W. (1989). Interpreting data from two-period crossover design. *Psychological Bulletin, 106*, 503-515.

Dansereau, D. F., Brooks, L. W., Holley, C. D., & Collins, K. W. (1983). Learning strategies training: Effects of sequencing. *Journal of Experimental Education, 51*, 102-108.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics, 21*, 467-480.

Hills, M., & Armitage, P. (1979). The two-period crossover clinical trial. *British Journal of Clinical Pharmacology, 8*, 7-20.

Jones, B., & Kenward, M. G. (1989). *Design and analysis of cross-over trials.* New York: Chapman and Hall.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.) (pp. 399-404). Englewood Cliffs, NJ: Prentice Hall.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Lentner, M., & Bishop, T. (1993). *Experimental design and analysis.* Blackburg, VA: Valley Book Company.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse Mental Rotations Test: Different versions and factors that affect performance. *Brain and Cognition, 28*, 39-58.

Ratkowsky, D. A., Evans, M. A., & Alldredge, J. R. (1993). *Cross-over experiments: Design, analysis, and application.* New York: Marcel Dekker.

Sensenig, L. D., Mazeika, E. J., & Topf B. (1989). Sign language facilitation of reading with students classified as trainable mentally-handicapped. *Education and Training of the Mentally Retarded, 24*, 121-125.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental Rotations: Group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47*, 599-604.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design.* New York: McGrawHill.

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◆

Please enter a subscription to Research in the Schools for:

Name: ____

Institution: ____

Address: ____

____

			COST
Individual Subscription ($25 per year)	Number of years	_____	_______
Institutional Subscription ($30 per year)	Number of years	_____	_______
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	_____	_______
TOTAL COST:			_______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
RESEARCH IN THE SCHOOLS
University of Alabama at Birmingham
School of Education, 233 Educ. Bldg.
901 13th Street, South
Birmingham, AL 35294-1250

Please note that a limited number of copies of Volume 1 are available and can be purchased for the same subscription prices noted above.

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

MSERA Membership/Renewal Form

(Please print or type)

Name ________________________________

Organization ________________________________

Address ________________________________

____

Telephone Work: ____________________

Home: ____________________

Fax: ____________________

e-mail: ______________________________

Amount Enclosed: MSERA 1997 Membership ($15 professional, $10 student) $________

MSER Foundation Contribution $________

TOTAL $________

Make check out to MSERA and mail to:

Dr. Gerald Halpin
Auburn University
4036 Haley Center
Auburn, AL 36849

RESEARCH IN THE SCHOOLS
Mid-South Educational Research Association
and the University of Alabama at Birmingham
901 South 13th Street, Room 233
Birmingham, AL 35294-1250

BULK RATE
U.S. POSTAGE
PAID
PERMIT NO. 1256
BIRMINGHAM, AL

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and the University of Alabama at Birmingham.

Volume 4, Number 2 **Fall 1997**

Racial Identity Attitudes and School Performance Among African American High School Students: An Exploratory Study 1
Steve R. Sandoval, Terry B. Gutkin, and Wendy C. Naumann

Evaluation of the Teaching Enhancements Affecting Minority Students (TEAMS) Program 9
Leanne Whiteside-Mansell, Nicola A. Conners, Melissa Crawford, and Richard Hanson

School Counselors' Perceptions of the Counseling Needs of Biracial Children in an Urban Educational Setting 17
Nancy J. Nishimura and Linda Bol

How Experienced Teachers Think About Their Teaching: Their Focus, Beliefs, and Types of Reflection 25
Rita M. Bean, Deborah Fulmer, Naomi Zigmond, and Judith V. Grumet

Teacher Perception of Kentucky Elementary Principal Leadership Effectiveness and School-Based Council Meeting Effectiveness 39
Patricia Lindauer, Garth Petrie, and Michael Richardson

Prevalence and Identification of Attention-Deficit Hyperactivity Disorder in a Mid-Southern State 49
Christine E. Daley, Harold Griffin, and Anthony J. Onwuegbuzie

A Case Study of an In-School Suspension Program in a Rural High School Setting 57
Tammye Turpin and Dawn T. Hardin

Score Comparisons of ACCUPLACER (Computer-Adaptive) and COMPANION (Paper) Reading Tests: Empirical Validation and School Policy 65
Jason C. Cole and Anthony D. Lutkus

James E. McLean and Alan S. Kaufman, Editors

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250. Please direct questions to jmclean@uab.edu**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1997 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 4, Number 2 **Fall 1997**

EDITORS
James E. McLean, *University of Alabama at Birmingham*
and Alan S. Kaufman, *Yale University School of Medicine*

PRODUCTION EDITOR
Margaret L. Rice, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Gypsy A. Abbott, *University of Alabama at Birmingham*
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
Larry G. Daniel, *The University of Southern Mississippi*
Paul B. deMesquita, *University of Rhode Island*
Donald F. DeMoulin, *University of Memphis*
R. Tony Eichelberger, *University of Pittsburgh*
Daniel Fasko, Jr., *Morehead State University*
Ann T. Georgian, *Hattiesburg (Mississippi) High School*
Tracy Goodson-Espy, *University of North Alabama*
Glennelle Halpin, *Auburn University*
Marie Somers Hill, *East Tennessee State University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
JinGyu Kim, *National Board of Educational Evaluation (Korea)*
Jwa K. Kim, *Middle Tennessee State University*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Jerry G. Mathews, *Idaho State University*
Peter C. Melchers, *University of Cologne (Germany)*
Claire Meljac, *Unité de Psychopathologie de l'Adolescent (France)*
Soo-Back Moon, *Catholic University of Hyosung (Korea)*
Arnold J. Moore, *Mississippi State University*
Thomas D. Oakland, *University of Florida*
William Watson Purkey, *The University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Georgia Southern University*
James R. Sanders, *Western Michigan University*
Anthony J. Scheffler, *Northwestern State University*
John R. Slate, *Valdosta State University*
Scott W. Snyder, *University of Alabama at Birmingham*
Bruce Thompson, *Texas A & M University*

GRADUATE STUDENT EDITORIAL BOARD
Margery E. Arnold, *Texas A & M University*
Vicki Benson, *The University of Alabama*
Alan Brue, *University of Florida*
Sue E. Castleberry, *Arkansas State University*
James Ernest, *University of Alabama at Birmingham*
Robin A. Groves, *Auburn University*
Harrison D. Kane, *University of Florida*
James C. Kaufman, *Yale University*
Sadegh Nashat, *Unité de Psychopathologie de l'Adolescent (France)*
Michael D. Scraper, *Kansas Newman College*
Sherry Vidal, *Texas A & M University*

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 1-8

Racial Identity Attitudes and School Performance Among African American High School Students: An Exploratory Study

Steve R. Sandoval, Terry B. Gutkin, and Wendy C. Naumann
University of Nebraska-Lincoln

This study investigated the relationship between racial identity attitudes and academic achievement among African-American adolescents. African-American students from a predominantly White midwestern city participated. The Pre-Encounter, Encounter, Immersion/Emersion, and Internalization subscales of the Racial Identity Attitude Scale (Helms & Parham, 1990), assessing racial identity development among African Americans, were correlated with school achievement measures. Results indicated negative correlations between Pre-Encounter scores, and Reading and Global Composite scores on the California Achievement Test (CAT). Immersion/Emersion scores were negatively associated with mid-semester, semester, and cumulative GPA; Global Composite scores on the CAT; and attendance. Internalization scores, however, were related positively with Cumulative GPA. These results indicate that academic achievement may be related to racial identity attitudes. Educational implications and directions for future research are discussed.

Schools seem to be in a state of crisis regarding the education of African-American youth and other historically subordinated students of color residing in the United States. Low scholastic aptitude test scores, grade point averages, attendance statistics, and high dropout rates reflect this phenomenon (Washington, 1988).

A number of authors have attempted to explain the poor school performance of African-American students, often pointing out the cultural discontinuities between minority communities and the Euro-American school system (Boateng, 1990; Carter, 1990; Erickson, 1987; Fillmore, 1988; Foley, 1991; Hale-Benson, 1990; Irvine, 1990; Moll & Diaz, 1987; Steinberg, Dornbusch & Brown, 1992). Specifically, many African-American students experience dissonance from the pressure to conform with Euro-American culture in order to succeed in school, while simultaneously receiving pressure from the minority community to retain their racial identity and avoid "acting White." For example, while doing homework has been found repeatedly to be associated strongly with school success (Keith & Benson, 1992; Mercure, 1993; Tymms & Fitz-Gibbon, 1992), it may be perceived by African-American peer groups as "selling out." The sense of past generations, that the success of one African-American person meant success for all African-Americans, seems to have been replaced among many African Americans with the perception that successful minorities have sacrificed their "cultural-selves" for personal gain (Fordham, 1988). As such, many African-American students find themselves caught "between a rock and a hard place," with the cultural expectations of their racial peers often standing in conflict with those of the dominant society. Lindstrom and San Vant (1986), who studied African-American gifted students, captured this dilemma in the statement of one such student who said, "I had to fight to be gifted and then I had to fight because I am gifted" (p. 584).

Fordham and Ogbu (1986) and Fordham (1988, 1991) report that African-American students have developed a variety of dissonance reducing mechanisms for coping with the conflicting messages and pressures of African-American and Euro-American cultures. In their qualitative work, Fordham and Ogbu found some students who openly admitted that being known as a "brainiac" (p. 187) (i.e., a high achieving student) had negatively affected their academic effort. Their dissonance was resolved by underachieving in school. Conversely, other students dealt with this conflict by totally accepting "the dominant Euro-American ideology that equates school-learning with the essence of civilization and progress" (Fordham, 1991, p. 474). In cases such as these, African-

---

We gratefully acknowledge the assistance of Elaine Werth and Thomas Christie for their helpful feedback on this article. Steve Sandoval received his Ph.D. in School Psychology this past summer from the University of Nebraska-Lincoln and is now a school psychologist with the Greeley Public Schools in Colorado. Terry Gutkin is a Professor of Educational Psychology and Director of the School Psychology Program at the University of Nebraska-Lincoln. Wendy Naumann is a doctoral student in School Psychology with a second major in Quantitative and Qualitative Methods at the University of Nebraska-Lincoln. Requests for reprints should be sent to Terry B. Gutkin, Department of Educational Psychology, 117 Bancroft Hall, University of Nebraska-Lincoln, Lincoln, NE 68588-0345 or emailed to tgutkin@unlinfo.unl.edu.

American adolescents had assumed an un-Black or "raceless" (Fordham, 1991, p. 480) persona for a higher value they had attached to school-related skills and credentials. Finally, according to Fordham and Ogbu, some high achieving African-American students managed their cultural dissonance by hiding and camouflaging their skills and competencies, thus avoiding the problems they might eventually encounter from their ethnic community.

In light of the cultural dissonance experienced by many African-American adolescents it would not be surprising if their academic achievement was affected substantially by their racial identity attitudes, that is, the way in which they view themselves as "racial beings." There is already a large literature base indicating that the self-concept and self-esteem of students in general (Kinney & Miller, 1988), and African-American students in particular (Mboya, 1986), are related to academic achievement. Further, although not considered a conclusive finding (Phinney, 1991), Paul and Fischer (1980) reported a positive relationship between racial identity and self-concept among African-American adolescents. Like self-concept and self-esteem, racial identity attitudes reflect basic self-perceptions held by individuals about themselves and their relationship to the world around them, and thus may also impact adolescents' school achievement. Beyond this, the manner in which adolescents understand their own racial identity might make them more or less vulnerable to the cultural dissonance that many African-American youth experience.

According to Ponterotto (1988), the most accepted theory of racial identity attitude development for African Americans is the one proposed by Cross (1971). He postulated four (originally five) different statuses of racial identity attitudes: Pre-Encounter, Encounter, Immersion/Emersion, and Internalization. The *Pre-Encounter* status characterizes individuals who have developed a worldview dominated by a Euro-American frame of reference and who negate or devalue their own "Blackness." The *Encounter* status characterizes African Americans who are beginning to question their identity because of experiencing an incident or event inconsistent with their original frame of reference. These persons may have experienced racial discrimination and, as a result, begun consciously to develop a Black identity. The *Immersion/Emersion* status involves African Americans who have acquired a high level of "Black pride." They struggle to rid themselves of anything considered "White" and cling to all elements of Blackness. African Americans who have reached the *Internalization* status are ones who have achieved a sense of satisfaction, inner security, and self-confidence with their Blackness. No longer are they "anti-White" and "pro-Black;" instead, they often become more pluralistic and seek companionship with people regardless of their race.

Considered the first major model of racial identity attitude development, Cross' (1971) initial work was integral in sparking other scholars to develop instruments for measuring racial identity attitudes of African Americans. Helms and Parham (1985, 1990) have perhaps been most ambitious in taking from Cross' original model to formulate a scale, namely the Racial Identity Attitude Scale (RIAS), utilized to assess where African-American individuals stand on each of Cross' four statuses.

In summary, there appears to be evidence suggesting a possible association between racial identity and school performance. In order better to understand this relationship, this study investigated whether the school achievement of African-American high school students was mediated by their racial identity, as measured by the RIAS (Helms & Parham, 1990).

Method

Participants

The sample consisted of 26 10th, 11th, and 12th grade African-American students from a predominantly White, medium size, midwestern city, who volunteered to participate in this study. There were approximately equal numbers of males and females in the sample (11 and 15, respectively). Participants ranged in age from 15 to 18 years (mean = 16 years, 8 months). All students were solicited from two of the city's four high schools that contained the highest percentage of African-American students. Only students who identified themselves as African-American students (including one racially-mixed student) were eligible to participate.

Dependent Variables

Academic achievement was assessed in a number of ways. First, participants' mid-semester, semester, and cumulative Grade Point Averages (GPAs), based on a 4.00 scale, were calculated. Mid-semester GPAs reflected participants' grades at the time their racial identity attitude was measured for the purposes of this study. Semester GPAs indicated students' grades at the completion of that semester. Cumulative GPAs measured the grades earned by participants throughout their years in high school. Since sophomores (*n* =12) were in high school for less than a year, it was not possible to calculate a Cumulative GPA for this group. These GPA data were complemented by standardized tests scores (Mathematics, Reading, and Global Composite) from the California Achievement Test (CAT) (1985). Participants' CAT scores were used only if the CAT was taken no more than six months prior to the study. As such, CAT scores from the seniors (*n* =5) were not included because this time criterion had not been met. In order to best interpret scores on the CAT, the normal curve equivalent (NCE) (*M* = 50, *SD* = 21) score was used. Finally, current attendance records were acquired.

Independent Variables

Racial Identity Attitude Scale (RIAS). The long-form of the Racial Identity Attitude Scale (RIAS-L) (Helms & Parham, 1990) is a 50-item scale assessing the four statuses (i.e., Pre-Encounter, Encounter, Immersion/Emersion, and Internalization) of Black identity development as described by Cross (1971). Using a 5-point Likert scale, participants indicated their identity attitude by marking whether they strongly disagreed (1) or strongly agreed (5) with each item. Sample items included the following: (a) "I believe that White people are intellectually superior to Blacks" [Pre-encounter], (b) "I am determined to find my Black identity" [Encounter], (c) "White people can't be trusted" [Immersion/Emersion], and (d) "A person's race has little to do with whether or not he or she is a good person" [Internalization].

The RIAS has been used extensively in prior research (Pope-Davis, Menefee, & Ottavi, 1993). Helms and Parham (1985) have reported that the Pre-Encounter, Encounter, Immersion/Emersion, and Internalization subscales for the RIAS-L have internal consistency reliability coefficients of .76, .51, .69, and .80, respectively.

Ponterotto and Wise (1987) conducted a validity study involving 186 African-American college students. They factor analyzed the original 30-item RIAS using an oblique rotation and found that the factor structure had supported the theoretical constructs in Cross' Pre-Encounter, Immersion/Emersion, and Internalization subscales. Little support, however, was found for the Encounter subscale. Additionally, concurrent validity evidence for the RIAS was found in its significant correlations with Milliones' (1980) Developmental Inventory of Black Consciousness (DIB-C) (Grace, 1984).

Demographic Data. In addition to the RIAS-L, students completed a Student Information Sheet providing a variety of demographic data (e.g., date of birth, gender, grade level) and other potentially useful information (e.g., educational attainment of parents, future plans after high school, and reactions to the study itself). Lastly, participants were asked to indicate (i.e., "yes-no") whether they had any difficulty reading the RIAS-L.

Procedures

Prior to the study, participants were informed that all data collection would be conducted in a manner that protected their anonymity and privacy. After the preliminary instructions were read and questions from the participants were answered, the students completed the RIAS-L and subsequently filled out the Student Information Sheet. GPAs, standardized achievement test scores, and attendance records were obtained from the public school district central office.

Results

Descriptive data from the Student Information Sheet are presented in Table 1. One hundred percent of the participants indicated having no difficulty reading the RIAS-L. Means and standard deviations for academic achievement measures and the RIAS are shown in Table 2.

Cronbach alpha reliability coefficients for the Pre-Encounter, Encounter, Immersion/Emersion, and Internalization subscales for this study were .67, .10, .77, and .32, respectively. Because the alpha reliability for the Encounter subscale was so low, no subsequent analyses utilizing this subscale were conducted. Correlations among the remaining three RIAS-L subscales indicated that each was relatively independent of the other, with the exception of a significant negative relationship between the Pre-Encounter and the Internalization subscales (see Table 3).

Table 1
Data from the Student Information Sheet

	Frequency	Percentage
Gender		
Male	11	42
Female	15	58
Mother's Education		
Some High School	05	19
High School Graduate	03	12
Some College	08	31
Received Bachelors Degree	05	19
Received Graduate Degree	04	15
Unknown	01	04
Father's Education		
Some High School	06	23
High School Graduate	03	12
Some College	04	15
Received Bachelors Degree	02	08
Received Graduate Degree	04	15
Unknown	07	27
Plans After High School		
Work Outside Home	01	04
Work Inside Home	00	00
Junior (2-year) College	02	08
Vocational/Technical School	00	00
Four-Year College/University	23	88
Not Sure	00	00

Table 2
Means and Standard Deviations of Academic Achievement Measures and the RIAS-L Subscale Scores

Variable	*n*	*Mean*	*SD*
Academic Achievement			
Mid-Semester GPA	26	2.36	0.80
Semester GPA	26	2.40	0.74
Cumulative GPA	14	2.53	0.73
NCE Reading	21	49.1	20.7
NCE Mathematics	21	45.8	19.5
NCE Global Composite	21	44.3	20.1
Percent Attendance	26	94.0	6.1
RIAS-L Subscales			
Pre-Encounter	26	2.01	0.45
Encounter	26	3.36	0.68
Immersion/Emersion	26	2.93	0.67
Internalization	26	4.14	0.31

Table 3
Subscale Intercorrelations

	RIAS-L Subscale	
RIAS-L Subscale	PE	I/E
Pre-Encounter		
Immersion/Emersion	.26	
Internalization	-.50**	-.08

Note. PE=Pre-Encounter, I/E=Immersion/Emersion
**$p<.01$

The principle analyses for this study were correlations conducted between the RIAS-L subscales and the various measures of academic achievement. Several statistically significant negative correlations were found for the Pre-Encounter and Immersion/Emersion subscales, while one statistically significant positive correlation emerged for the Internalization subscale (see Table 4).

Table 4
Intercorrelations Between Racial Identity and Academic Achievement

		RIAS-L Subscale		
Academic Achievement	*n*	PE	I/E	I
Mid-Semester GPA	26	-.34	-.72**	.19
Semester GPA	26	-.22	-.40*	.23
Cumulative GPA	14	-.46	-.63*	.73*
Reading (in NCE units)	21	-.52*	-.41	.29
Mathematics (in NCE units)	21	-.24	-.32	.35
Global Composite (in NCE units)	21	-.47*	-.44*	.42
Attendance	26	-.12	-.54**	-.15

Note. PE=Pre-Encounter, I/E=Immersion/Emersion, I=Internalization
*$p<.05$., **$p<.01$.

Discussion

This study investigated the relationship between racial identity attitudes and academic performance among African-American high school students. Results indicated that academic achievement was negatively related to both Pre-Encounter and Immersion/Emersion attitudes, and on one measure of academic performance (cumulative GPA), a positive association was found with Internalization attitudes. Given the correlational nature of this study, it is, of course, not possible to draw any firm conclusions from these data regarding causal relationships. Acknowledging this at the outset, the following discussion of the results presents a number of plausible explanations that are worthy of future research and investigation.

Possible Impact of Racial Identity Attitudes on Achievement

According to several "cultural conflict" theorists (Delgado-Gaitan, 1988; Moll & Diaz, 1987; Patthey-Chavez, 1993; Trueba, 1987), there is a widely held belief that among students of color in the United States, total assimilation to "Euro-American" culture is the "best" way to achieve successfully both in school and society in general. The findings from this study, however, are not completely consistent with this view. For instance, although the Immersion/Emersion results indicate that students who "reject" White culture are not associated with having high achievement scores, analyses of the Pre-Encounter data run counter to the "assimilationist" perspective. Specifically, African-American students who tended to assimilate to a Euro-American worldview did not show higher patterns of academic success. Perhaps even more importantly, students who adhered to a more bicultural, pluralistic, and non-prejudiced worldview (i.e., those with Internalization attitudes) were able to maintain a strong cultural identity while achieving academic success at the same time.

Pre-Encounter Attitudes. Regarding the Pre-Encounter results, the negative correlations found for the Reading and Global Composite scores of the CAT were not consistent with Fordham's (1988, 1991) and Fordham and Ogbu's (1986) findings that being "un-black" was associated with higher academic achievement. On the contrary, these findings revealed a negative relationship between the extent to which African-American students adopted a White worldview and their academic achievement. Since Pre-Encounter attitudes reflect negative evaluations of the African-American racial/ethnic group, one possible explanation for this result may be that many of these students expected little of their own academic capabilities due to their membership in what they may have considered an inferior group. Empirical studies that have shown a relationship between low self-expectations and low performance in school (Alderman & Doverspike, 1988; Holahan, 1981; House, 1993) are consistent with this explanation.

In addition, research has revealed that experiencing high stress is inversely associated with academic achievement (Grannis, 1992; Hackett, Betz, Casas, & Rocha-Singh, 1992; Matthews & Burnett, 1989). Because many students who scored high on the Pre-Encounter subscale are more likely to have developed negative attitudes toward their own racial/ethnic group, they are perhaps less likely to receive support and peer affiliation from their own cultural group. Instead, they may be more likely to be on the receiving end of hostile remarks and/or behaviors from their peers of the same race (Fordham & Ogbu, 1986), perhaps leading to higher levels of stress compared with other students. These higher levels of anxiety experienced by many students with high Pre-Encounter scores may be linked partially to their lower academic scores.

Moreover, academic achievement and self-esteem have been shown to be related strongly (Kinney & Miller, 1988). This relationship appears to be particularly true among African-American students (Mboya, 1986). According to Parham and Helms (1985), many students scoring high on the Pre-Encounter subscale also show patterns of low self-esteem levels. This finding may indicate that for many such students, their academic performance may have been influenced negatively by their low self-concepts.

Immersion/Emersion Attitudes. Participants' scores on the Immersion/Emersion subscale were negatively related to mid-semester, semester, and cumulative GPAs; the Global Composite score on the CAT; and attendance. That is, the stronger the attachment of participants to their own cultural "blackness," the lower their performance on numerous academic achievement measures. These findings may be explained by *cultural inversion* (Ogbu, 1987). That is, many African Americans have learned to devalue the cultural worldview of the dominant group as a means of protecting their own cultural interests and identity. According to Fordham and Ogbu (1986), historically, African Americans were denied acknowledgment of their intellectual capabilities by the majority group and thus many African Americans began to denigrate themselves along this dimension and regard academic achievement as a White person's prerogative. As such, African-American students who strive for academic success may appear to many of their peers to be "acting White." Since Immersion/Emersion students are least likely to exhibit behaviors they perceive as "White," it is understandable that their academic indicators would be negatively related to their Immersion/Emersion attitudes.

Also, like the pattern of low self-esteem found among many African-American students scoring high on the Pre-Encounter subscale, similar results have been found among students scoring high on the Immersion/Emersion subscale (Parham & Helms, 1985). As such, the strong relationship between self-esteem and academic achievement (Kinney & Miller, 1988) may help explain further the prevalence of low achievement scores among many students scoring high on the Immersion/Emersion subscale.

Internalization Attitudes. Participants' scores on the Internalization subscale were positively and strongly associated with cumulative GPA. A possible explanation for this intriguing finding may be that students with high Internalization scores had resolved their cultural dissonance to the extent that it did not impact negatively their overall school performance. Cross (1971) had conceptualized the Internalized individual as one who was more secure and self-confident with his or her sense of Self as a racial being. Consequently, students with high Internalization scores may feel less compelled to perceive "successful" academic behaviors as those only White persons should obtain. Support for this explanation comes from a number of studies showing a trend for Internalization and closely related attitudes to be associated positively with high self-esteem (Parham & Helms, 1985; Phinney & Alipuria, 1990; Phinney, Williamson, & Chavira, 1990).

The interpretation of the positive and strong correlation between internalization attitudes and cumulative GPA must be treated with some caution, however. Specifically, although most of the other six correlations with academic achievement were positive and many were at least moderate in magnitude, they did not achieve statistical significance.

Possible Impact of Achievement on Racial Identity Attitudes

While the analyses discussed thus far have all focused on how racial identity attitudes may have impacted school achievement, given the correlational nature of the study it is equally plausible that the reverse also may have been true. That is, students' levels of academic achievement may have impacted the development of racial identity attitudes. It seems logical, for example, that poor school performance may have increased Pre-Encounter attitudes among some students by reinforcing perceptions that their own racial/ethnic group was, in fact, inferior to Whites. Other students with poor academic records, however, may have rejected this notion and chosen instead to attribute their problems to the academic environment which they perceive as not meeting their needs. Given the White, middle-class culture that predominated the schools included in this study, lower academic achievement may have led these students to a rejection of the dominant culture, a closer affiliation and alignment with their own cultural group, and thus an increase in their Immersion/Emersion attitudes. Finally, among those African-American students who performed well in school, the experience of academic success may have fostered their conceptions that "cultural" and "academic" success could be achieved simultaneously. That is, despite many African-American students having to cope with the pressures between their own cultural peer group and the school's "assimilation" agenda (Fordham & Ogbu, 1986), successful African-American students may have been able to see that achieving academically was possible without having to sacrifice their racial identity. This experience may have promoted higher Internalization attitudes among these students.

Possible Impact of Other Variables on Both Racial Identity Attitudes and Achievement

Finally, given the correlational nature of this study, it is possible that racial identity attitudes and academic achievement are not related directly to each other at all. That is, the significant correlations reported in this investigation may be a function of one or more variables (e.g., intelligence, socio-economic status) to which both racial identity attitudes and academic achievement are related.

Limitations of the Study and Directions for Future Research

Perhaps the most notable limitations of the study center around possible threats to its external validity. First and foremost, the sample size (N=26) was quite small. Second, since all participants were solicited from a predominantly Euro-American, midwestern city, the reader should be wary of generalizing these findings to the attitudes of African-American adolescents residing in different geographical locations with different racial mixes. Specifically, the immediate social context of the school may determine if one's ethnicity is a salient issue. In this study, the participants were all part of a small percentage of African Americans enrolled in their respective schools. Feeling marginalized as persons of "ethnic" heritage may have impacted their racial identity attitudes and the relationship of these attitudes to academic behavior. On the other hand, in schools that are largely African American (e.g., inner-city Chicago), ethnicity may not be a salient characteristic that African-American adolescents use to guide their interactions because they are not singled out as different. Future researchers should thus consider how racial identity attitudes might be influenced by social context.

Additionally, the sample employed in this study may not have been representative of the African-American students in the two high schools from which the sample was drawn. Specifically, according to the Student Information Sheet (see Table 1), all but three participants (88%) mentioned wanting to attend a four-year college or university upon high school graduation, and a high percentage of participants' mothers and fathers had attended or completed college. In addition, the participants' average cumulative GPA (2.53), and Reading, Mathematics, and Global CAT scores (49.1, 45.8, and 44.3, respectively) were substantially higher than the average cumulative GPA (1.98), and Reading, Mathematics, and Global CAT scores (43.1, 40.5, and 39.5, respectively) of African-American students residing in the city where the study was conducted. As such, the findings reported in this study may be reflective primarily of higher achieving African-American high school students.

Unfortunately, the alpha reliability coefficient for the Internalization subscale was quite low. The reader should, therefore, be particularly cautious when interpreting the findings associated with this subscale.

Finally, an important and unresolved question is the issue of causality. Because this study was correlational in nature, there was no way to determine whether: (a) racial identity attitudes affected academic achievement, (b) academic achievement affected racial identity, (c) the relationship between racial identity and academic achievement was bi-directional, or (d) there was no causal connection between racial identity and academic achievement. If possible, future researchers should manipulate experimentally racial identity attitudes and academic achievement to determine the directionality of influence that each exerts on the other.

Educational Implications

Since this study has produced some evidence that the development of Internalization attitudes may relate to some forms of increased academic achievement and virtually everyone would prefer that all adolescents espouse non-prejudiced beliefs such as those associated with Internalization attitudes, school psychologists and other educators should begin to consider how they might create school environments that promote Internalization attitudes among African-American youth. Unfortunately, it is unlikely that this worldview will develop simply by happenstance. According to Helms (1990), Internalization attitudes often result from interacting with a mainstream member of society who has earned mutual trust, respect, and acceptance. Given the large amount and intensity of time children and adolescents spend in schools, however, school psychologists and other educators would seem to be in a prime position to establish such relationships with their African-American students, thus helping these students to develop Internalization attitudes.

Clearly, one positive step would be to develop schools which remove the perception among African-American students that academic achievement can only be accomplished by "acting White." Building on research pertaining to alternative pedagogies, instruction, curricula, and teacher behaviors (Irvine, 1990; Justiz & Darling, 1980; Slavin, 1985; Sleeter & Grant, 1988), school personnel must find ways for their African-American students to feel "culturally safe" while attaining their full academic potential. It is unlikely that either Internalization attitudes or academic achievement will flourish among African-American students as long as they perceive that they must choose between "being educated" and "being Black." As organizational (Schmuck, 1990; Snapp, Hickman, & Conoley, 1990) and case consultants (Gutkin & Curtis, 1990), school psychologists could play a pivotal role in helping educators establish "culturally safe" school environments.

References

Alderman, M. K., & Doverspike, J. E. (1988). Perceived competence, self-description, expectation, and successful experience differences among students in grades seven, eight, and nine. *Journal of Early Adolescence, 8*, 119-131.

Boateng, F. (1990). Combating deculturalization of the African-American child in the public school system: A multicultural approach. In K. Lomotey (Ed.), *Going to school: The African-American experience* (pp. 209-222). Albany, NY: State University of New York Press.

California Achievement Test. (1985). Englewood-Cliffs, NJ: CTB/McGraw-Hill.

Carter, R. T. (1990). Culture and Black students' success. *Educational Considerations, 18*, 7-11.

Cross, W. E., Jr. (1971). The Negro-to-Black conversion experience: Toward a psychology of Black liberation. *Black World, 20*, 13-27.

Delgado-Gaitan, C. (1988). The value of conformity: Learning to stay in school. *Anthropology & Education Quarterly, 19*, 354-381.

Erickson, F. (1987). Transformation and school success: The politics and culture of educational achievement. *Anthropology and Education Quarterly, 18*, 335-356.

Fillmore, L. W. (1988, Summer). *Now or later? Issues related to the early education of minority group children.* Paper presented to the Council of Chief State School Officers summer meetings, Boston, MA.

Foley, D. E. (1991). Reconsidering anthropological explanations of ethnic school failure. *Anthropology and Education Quarterly, 22*, 61-86.

Fordham, S. (1988). Racelessness as a factor in Black students' success: Pragmatic strategy or pyrrhic victory? *Harvard Educational Review, 58*, 54-84.

Fordham, S. (1991). Racelessness in private schools: Should we deconstruct the racial and cultural identity of African-American adolescents? *Teachers College Record, 92*, 470-484.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting white.'" *The Urban Review, 18*, 176-206.

Grace, C. (1984). *The relationship between racial identity attitudes and choice of typical and atypical occupations among Black college students.* Unpublished doctoral dissertation. Teachers College, Columbia University. New York.

Grannis, J. C. (1992). Students' stress, distress, and achievement in an urban intermediate school. *Journal of Early Adolescence, 12*, 4-27.

Gutkin, T. B., & Curtis, M. J. (1990). School-based consultation: Theory, techniques, and research. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 577-611). New York: Wiley.

Hackett, G., Betz, N. E., Casas, J. M, & Rocha-Singh, I. A. (1992). Gender, ethnicity, and social cognitive factors predicting the academic achievement of students in engineering. *Journal of Counseling Psychology, 39*, 527-538.

Hale-Benson, J. (1990). Visions for children: Educating black children in the context of their culture. In K. Lomotey (Ed.), *Going to school: The African-American experience* (pp. 209-222). Albany, NY: State University of New York Press.

Helms, J. E. (1990). *Black and White racial identity: Theory, research, and practice.* Westport, Connecticut: Greenwood Press.

Helms, J. E., & Parham, T. A. (1985). *The Racial Identity Attitude Scale (RIAS).* Unpublished manuscript.

Helms, J. E., & Parham, T. A. (1990). Black racial identity attitudes scale (Form RIAS-B). In J. E. Helms (Ed.), *Black and White racial identity.* New York: Greenwood.

Holahan, C. K. (1981, August). *Student perceptions and social comparisons and performance expectancy.* Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

House, J. D. (1993). Achievement-related expectancies, academic self-concept, and mathematics performance of academically underprepared adolescent students. *Journal of Genetic Psychology, 154*, 61-71.

Irvine, J. J. (1990). *Black students and school failure: Policies, practices, and prescriptions.* New York: Greenwood Press.

Justiz, M. J., & Darling, D. W. (1980). A multicultural perspective in teacher education. *Educational Horizons, 58,* 203-205.

Keith, T. Z., & Benson, M. J. (1992). Effects of manipulable influences on high school grades across five ethnic groups. *Journal of Educational Research, 86,* 85-93.

Kinney, P., & Miller, M. J. (1988). The relationship between self-esteem and academic achievement. *College Student Journal, 22,* 358-362.

Lindstrom, R. R., & San Vant, S. (1986). Special issues in working with gifted minority adolescents. *Journal of Counseling and Development, 64,* 583-586.

Matthews, D. B., & Burnett, D. D. (1989). Anxiety: An achievement component. *Journal of Humanistic Education and Development, 27,* 122-131.

Mboya, M. M. (1986). Black adolescents: A descriptive study of their self-concepts and academic achievement. *Adolescence, 21,* 689-696.

Mercure, C. M. (1993). Project achievement: An after school success story. *Principal, 73,* 48-50.

Milliones, J. (1980). Construction of a Black consciousness measure: Psychotherapeutic implications. *Psychotherapy: Theory, Research, and Practice, 17,* 175-182.

Moll, L., & Diaz, S. (1987). Change as the goal of educational research. *Anthropology and Education Quarterly, 18,* 300-311.

Ogbu, J. U. (1987). Variability in minority school performance: A problem in search of an explanation. *Anthropology and Education Quarterly, 18,* 312-334.

Parham, T. A., & Helms, J. E. (1985). Relation of racial identity attitudes to self-actualization and affective states of Black students. *Journal of Counseling Psychology, 32,* 431-440.

Patthey-Chavez, G. G. (1993). High school as an arena for cultural conflict and acculturation for Latino Angelinos. *Anthropology and Education Quarterly, 24,* 33-60.

Paul, M., & Fischer, J. (1980). Correlates of self concept among Black early adolescents. *Journal of Youth and Adolescence, 9,* 34-49.

Phinney, J. S. (1991). Ethnic identity and self-esteem: A review and integration. *Hispanic Journal of Behavioral Sciences, 13,* 193-208.

Phinney, J., & Alipuria, L. (1990). Ethnic identity in college students from four ethnic groups. *Journal of Adolescence, 13,* 171-183.

Phinney, J., Williamson, L., & Chavira, V. (1990, April). *Attitudes towards integration, assimilation, and separation among high school and college students.* Paper presented at the Western Psychological Association Meeting, Los Angeles.

Ponterotto, J. G. (1988). Racial consciousness development among White counselor trainees: A stage model. *Journal of Multicultural Counseling and Development, 16,* 116-126.

Ponterotto, J. G., & Wise, S. L. (1987). Construct validity study of the racial identity attitude scale. *Journal of Counseling Psychology, 34,* 218-223.

Pope-Davis, D. B., Menefee, L. A., & Ottavi, T. M. (1993). The comparison of White racial identity attitudes among faculty and students: Implications for professional psychologists. *Professional Psychology: Research and Practice, 24,* 443-449.

Schmuck, R. A. (1990). Organization development in schools: Contemporary concepts and practices. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 899-919). New York: Wiley.

Slavin, R. E. (1985). Cooperative learning: Applying contact theory in desegregated schools. *Journal of Social Issues, 41,* 45-62.

Sleeter, C. E., & Grant, C. A. (1988). *Making choices for multicultural education: Five approaches to race, class, and gender.* Macmillan Publishing Company: New York.

Snapp, M., Hickman, J. A., & Conoley, J. C. (1990). Systems interventions in school settings: Case studies. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 922-934). New York: Wiley.

Steinberg, L., Dornbusch, S. M., & Brown, B. B. (1992). Ethnic differences in adolescent achievement: An ecological perspective. *American Psychologist, 47,* 723-729.

Trueba, H. T. (1987). *Success or failure? Learning and the Language Minority Student.* New York: Harper & Row (pp. 15-33).

Tymms, P. B., & Fitz-Gibbon, C. T. (1992). The relationship between part-time employment and A-level results. *Educational Research, 34,* 193-199.

Washington, E. D. (1988). A componential theory of culture and its implications for African-American identity. *Equity and Excellence, 24,* 24-30.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 9-16

Evaluation of the Teaching Enhancements Affecting Minority Students (TEAMS) Program

Leanne Whiteside-Mansell, Nicola A. Conners, Melissa Crawford, and Richard Hanson
University of Arkansas at Little Rock

Teaching Enhancements Affecting Minority Students (TEAMS) is a program designed to increase retention of minority students in higher education. Two intermediate goals of the TEAMS program are to increase minority students' knowledge about university services and participant satisfaction with the university experience. The major goal of the TEAMS program is to increase the institution's minority retention rates. A survey was mailed to all minority, US citizens enrolled as of Fall 1994 in a non-residential, urban institution of higher education with a majority of White students, where the TEAMS program had been implemented for three years. The survey assessed the minority students' knowledge about and use of a variety of student services and feelings about their educational experience. Logistic regression was performed controlling for student gender, GPA, and academic level. Retention rates for eight years were examined. Results of this study supported the continuation of the TEAMS program. More TEAMS members were aware of student services than non-TEAMS members. TEAMS members reported more positive views of their experiences in general than non-TEAMS members. Retention rates indicate a general upward trend.

Minority retention in higher education remains a challenge for universities today. Since 1980, the US Department of Education has been tracking the high school graduation class of that year. According to their findings, "normal persistence" in college is not the norm, especially for minority students. Of the class of 1980 minority students who enrolled in college full time, only one of seven African Americans continued full time for four years (Mingle, 1987). Stewart (1992) reported that in 1991 only 25% of minority students were graduating six years after entering college, compared to about half of white college students.

Low minority representation in graduate school is also a concern. While African-American students comprise about 13% of students who go to college, they constitute only 4.1% of graduate students. From 1975 to 1994, the number of doctorates awarded to African-American students has shown a slow increase from 3.8% to 4.1%, however, evidence in recent years indicates a decrease in the number of doctoral degrees. For example, the 1994 rate of 4.1% was down from 4.2% in 1993 and 4.4% in 1977 (Simmons & Thurgood, 1995).

Because one of the reasons that students leave school is related to the psychosocial climate of the educational setting, some research suggests that the key to retention is student involvement in campus activities (Adams, 1992). The more involved a student is in the social system of a university, the more likely he/she is to persist there. Social interaction with peers or faculty and participation in extracurricular activities are positively associated with persistence, degree attainment, and graduate school attendance. Involvement in campus life exposes students to other high achieving peers, reinforcing the academic goals of the student. Involvement also may help students meet their goals by facilitating personal development in areas such as interpersonal skills and self-confidence (Pascarella & Terenzini, 1991).

For minority students, there are often barriers to involvement in university life. In predominantly white institutions the faculty are traditionally white, and therefore minority students may have less access to social support. They may find the university environment alienating or even racist (Jacobi, 1991; Kobrak, 1992). Also, minority students are more likely to have attended inner city high schools or to be first generation college students, which may make their transition to college

---

Leanne Whiteside-Mansell is an Assistant Professor at the University of Arkansas at Little Rock and received her doctorate in Educational Research from the University of Memphis. Nicola A. Conners is a doctoral student at the University at Memphis and is currently part of the evaluation team of the Women and Children Recovery Center. Melissa Crawford has a masters in Interpersonal and Organizational Communication, was a past coordinator of the TEAMS program, and is the director of the Speech Communication Interactive Learning Center at the University of Arkansas at Little Rock. Richard Hanson is the Dean of the University of Arkansas at Little Rock Graduate School. We gratefully acknowledge the assistance of Hallethia Wofford and Carmelita King. Correspondence should be directed to Leanne Whiteside-Mansell, Center for Research on Teaching and Learning, University of Arkansas at Little Rock, 2801 S. University Avenue, Little Rock, AR 72204 or by e-mail at lawhiteside@ualr.edu.

difficult (Jacobi, 1991). African-American students, like low income college students in general, may find the transition difficult, because the university atmosphere is different socially, academically, and even culturally (Terenzizi et al., 1994).

One way to increase a student's involvement in the university community and avoid feelings of alienation is through mentoring (Adams, 1993). Mentoring involves the student with at least one peer or faculty member and gives him/her a greater tie to the university. Many universities are implementing mentoring programs for at-risk students (Jacobi, 1991). The belief is that by providing academic and social support, mentoring can improve achievement, increase retention, and feed the pipeline to graduate school. Since students leave school for reasons other than academic ones, effective mentoring involves more than academic support alone (Lewis, 1986; Redmond, 1990). Other causes of student attrition must be addressed, including the lack of knowledge about or access to social or academic resources, and the lack of psychological comfort with the university atmosphere (Redmond, 1990). Mentoring is important not only to decrease attrition, but also to encourage students to attend graduate school. Research suggests that personal faculty encouragement is a very important influence on a student's decision to attend graduate school (Pascarella & Terenzini, 1991). Mentoring could be especially helpful for minority students, because African-American students at pre-dominantly white colleges are far less likely than White students to seek counseling or tutoring on their own. Some suggest that this is due to the lack of African-American faculty or advisors at most universities (Wiley, 1989). The literature suggests that African-American students greatly prefer African-American mentors or role models, but they often are not available (Wiley, 1989; Jacobi, 1991). So, while minority students are often at-risk and have the greatest need for a role model, one is not often available. Planned mentoring would meet that need.

This study examines the effectiveness of a mentoring based program intended to increase the retention rate of minority students in a predominantly white, non-residential, state-supported, urban university. Although the evidence suggests that mentoring programs are helpful, the majority of research has been directed toward the traditional university setting (Lewis, 1986). The effectiveness of mentoring programs in the urban commuter college is less clear. Because of the diverse populations served by different types of institutions, others have argued for institution-specific and student-specific research (Peterson, 1993).

This study examined two areas that are thought to impact student retention -- services and satisfaction. Three evaluation questions were addressed. Are TEAMS students more aware than other minority students of services provided by the university? Are TEAMS students more satisfied with their experiences in the university setting than other minority students? Has the TEAMS program had an impact on minority student retention?

Method

Program Description

The Teaching Enhancements Affecting Minority Students (TEAMS) program is intended to increase retention by increasing the knowledge of minority students about services available on campus and their satisfaction with the college experience. The TEAMS program is directed to any minority student including African American, Hispanic, Asian, and American Indian. Goals of the program also include recruitment and encouragement of minority students to earn advanced degrees, however, this study will focus on the goal of retention. Graduate students, upperclassmen, staff, and faculty serve as volunteer mentors. The mentors are called upon to provide not only social interaction, but also summer research experience.

Three strategies are used to recruit students into the TEAMS program. First, letters are sent to all minority students at the beginning of the Fall Semester describing the program and inviting students to join. Second, TEAMS Graduate Assistants recruit students at all undergraduate orientation programs. Finally, during the first weeks of the fall semester, a TEAMS rally is held with minority Greek organization participation.

The TEAMS program involves students in many activities designed to orient them to University life. Students are invited to workshops covering topics such as study skills or time management, bi-weekly meetings of TEAMS groups, tutoring, speaking events, and summer research programs. TEAMS also sponsors many activities to help students become more involved in their school and community. Students may participate in luncheons, holiday parties, organized volunteer efforts, and outings to ball games or the theater. TEAMS students provide feedback to the program on a regular basis to select topics for meetings and plan events.

Students participating in TEAMS are divided into three groups. T-1 teams are groups of 5-10 freshmen and sophomores that are led by an upper-class mentor, a staff mentor, and a faculty mentor. These teams focus on supporting basic skills, building confidence, and providing direction for the students. Members of T-1 teams attend weekly meetings as well as tutoring sessions. This level has had the most student involvement. For example, 78 freshmen and sophomores were involved at this level during the second year of the program.

T-2 teams are formed around areas of professional interests (education, business, etc.) and usually consist of juniors and seniors. Student involvement is generally lower at this level. For example, 46 students were active T-2 members during the second year of the program. These teams focus less on survival skills and more on the professional development of students. To accomplish this, African Americans from the community are invited to speak about their professions. Additionally, T-2 members attend and present papers at professional meetings and work closely with faculty mentors on summer research experience. Funds are available for eight upper-class undergraduate TEAMS students to receive a stipend for a summer research experience.

T-3 teams are designed to establish one-on-one mentoring relationships between faculty members and students. These students are encouraged to work toward terminal degrees in their field. They may receive expense money to travel to conferences or other graduate schools. It is hoped that many of these students will rejoin the faculty at the completion of a terminal degree. For these students, TEAMS supports doctoral fellowships for the years they attend another university. At the end of the second year of this program 43 students were active at this level.

Mentors are recruited from the university faculty and research staff, and TEAMS students have input in the selection of mentors. For example, TEAMS students interested in summer research experience interview prospective mentors to participate in the selection of their summer placement.

Subjects

Data were collected from students at a predominantly white, nonresidential, state-supported, metropolitan university where the TEAMS program had been in effect for three years. The university had a student population of about 12,000 students and 400 full-time faculty in the 93-94 academic year. The university setting is non-traditional with many older (56% between 22 and 39 years of age), part-time (46%), and female (58%) students. It is not surprising to find that 85% of the students work 20 or more hours per week and have family responsibilities. A majority of students classify themselves as White (78%). African Americans comprise the majority of the minority population, however 26% of the minority population are Hispanic, Asian, American Indian, or some other minority. About 35 minority faculty members are in tenure track positions. The university offers 50 undergraduate major programs, 28 graduate and professional programs, and three doctoral programs besides the juris doctor.

Identification of TEAMS Members

TEAMS students were identified in two ways. For the two evaluation questions concerning knowledge of service and satisfaction with university life, students identified themselves. For the examination of retention rates of TEAMS students, all students that had enrolled in the TEAMS program were considered TEAMS members. The two methods were used for practical as well as confidentiality reasons. It was not practical to pre-mark surveys or pre-identify TEAMS members before mailing the survey. In addition, coding that identified students would violate the confidentiality of the survey. Assuring confidentiality was thought to be an important factor in both the response rate and accuracy of responses received. Neither of these methods account for the level of student participation in TEAMS. The TEAMS program is designed to allow the students to determine their own involvement in the program. Therefore this evaluation attempted to determine if program impact could be detected regardless of the level of student participation.

Service Use and University Satisfaction Survey

The effectiveness of the TEAMS program to increase knowledge of services and satisfaction was evaluated using a written survey instrument. A survey was mailed to all minority, US citizens enrolled at the institution as of Fall 1994, three years after the TEAMS program began. The survey consisted of three sections: Demographic and self-reported academic information (Table 1), a list of student services (Table 2), and statements concerning attitude and perception toward the university and university community (Table 3). The total number of surveys mailed (N = 2100) included all levels – undergraduates, graduates, and law students. Forty surveys were returned for lack of a valid address. The surveys were mailed at the end of the fall semester with a return date request of December 31. Of the 459 (22%) responses received, 75 (16.7% of responses) were TEAMS members and 384 (83.3% of responses) were non-TEAMS members.

Table 1 describes the survey sample. Respondents were identified as TEAMS members by their response on the survey. Respondents were asked if they were aware of 10 services available at the institution; TEAMS was one of these services. The level and current activity of TEAMS involvement was not assessed since students identified themselves as using or having used TEAMS in the past. That is, students were grouped as TEAMS members if they reported *ever* being involved in the TEAMS program. Respondents that indicated they "use or have used this service" regardless of their level of satisfaction were regarded as TEAMS members. With the exception of gender of the respondent, TEAMS respondents were similar to the non-TEAMS respondents on personal (age, hours worked during school, sources of funding) and academic characteristics (student-reported GPA, academic level). A higher percentage of the TEAMS respondents were female (82.67%) than the non-TEAMS respondents (69.79%). However, the high percentage of female respondents is not inconsistent with the student population nor of the population of TEAM members. Sixty-eight percent of African American students at the institution are female, and 72% of TEAM members are female.

Table 1
Description of Minority Student Respondents on Personal and Academic Characteristics

	TEAM member		Non-TEAM member	
Characteristic	*N*	Percent	*N*	Percent
Total Respondents	75		374	
Female	62	82.67	261	69.79
Age in Years				
< 24	40	53.33	165	44.24
25-35	14	18.67	113	30.29
35 +	21	28.00	95	25.47
Academic Level				
Freshman or Sophomore	24	32.00	175	47.95
Junior or Senior	24	32.00	107	29.32
Graduate	27	36.00	83	22.73
Hours work during Fall '94				
0 - 20	32	43.25	126	33.87
21-40	25	33.78	157	42.21
40 +	17	22.97	89	23.92
Source of funding- First Undergraduate Semester				
Self	14	18.92	94	25.47
Family	6	8.11	41	11.11
Loans	6	8.11	41	11.11
Grants/Scholarships	29	39.19	128	34.69
Other (or multiple)	9	25.67	65	17.62
Source funding - Fall 1994				
Self	20	27.78	126	34.81
Family	4	5.56	26	7.18
Loans	15	20.83	62	17.13
Grants	24	33.33	88	24.31
Other (or multiple)	9	12.50	60	16.57
	N	*Mean*	*N*	*Mean*
Average GPA	72	3.04	348	2.90

Table 2
Number and Percent of Minority Students That are Aware of Services

	TEAMS member		Non-TEAMS member	
Services	*N*	Percent	N	Percent
Student Financial Aid Services	66	92.96	339	91.13
New Student Orientation	56	76.71	279	75.20
Counseling and Career Planning	57	79.17+	248	66.67
Writing Lab	54	73.97	252	68.11
Math Lab	53	73.61	260	69.89
Academic Advising	68*	94.44	339	91.37
Campus Bookstore	74	98.67	364	98.11
Student Support Services	44	61.97*	144	38.61
Library Services	73	97.33	353	94.39

+significant before correction for number of tests
*significant after correction for number of tests

Compared to all TEAMS members, the respondents that identified themselves as TEAMS members were similar in academic level. Of the total active TEAMS members, 78 (46.7%) were freshmen or sophomores, 46 (27.6%) were juniors or seniors, and 43 (25.7%) were graduates. As shown in Table 1, TEAMS respondents had a somewhat similar distribution, however, respondents were slightly more often freshmen or sophomores.

In an attempt to statistically control for factors other than TEAMS that might impact use of services and attitudes, analyses of survey data included controls for gender, student GPA, and academic classification level.

Retention Data

In addition to the self-report survey, enrollment records for TEAMS members were examined as well as enrollment and retention of minority students at the institution. All students involved with TEAMS and enrolled during the fall semester of 1994 were tracked for four semesters (Fall 1994, Spring 1995, Fall 1995, Spring 1996). Sources of data include internally published reports on retention from the university's Office of Institutional Research and Budget and enrollment records for individual TEAMS students.

Results

Service Use

Respondents were asked to indicate if they were aware of each service. Differences between the two groups (TEAMS and non-TEAMS) for each service and attitude statement were examined using Chi-square. Because the likelihood of chance significant findings is great when numerous analyses are conducted, Bonferroni corrections were used. Table 2 shows the number and percent of TEAMS and non-TEAMS respondents indicating that they were aware of the services. Most of the services listed in Table 2 are self-explanatory with the possible exception of Student Support Services. Student Support Services is a free program designed to help students who need remediation. Services include tutoring, guidance, and counseling.

The majority of both groups indicated that they were aware of Student Financial Aid Services, Academic Advising, the Bookstore, and Library Services. Three other services (New Student Orientation, Writing Lab and Math Lab) were known by about 75% of the respondents, regardless of TEAMS membership. More TEAMS members (79.17%) indicated that they were aware of Counseling and Career Planning than non-TEAMS members (66.67%), however, this difference was not statistically significant after corrections were made for multiple tests. Significantly more TEAMS respondents (61.9%) were aware of the Student Support Services compared to (38.61%) non-TEAMS respondents.

Multivariate statistical analyses (logistic regression) were performed to control for gender, academic level, and student reported GPA. Logistic regression, like ordinary least squares regression, controls for other independent variables in the model, however, unlike ordinary least squares regression, logistic regression treats the dependent variable as a probability value and is appropriate when the dependent variable is dichotomous. Using logistic regression, the effect of the independent variable (TEAM membership) can be evaluated as substantively important by interpretation of the odds ratio. After controlling for gender, academic level, and student-reported GPA, TEAMS members were found to be 1.7 times more likely to be aware of Student Support Services than non-TEAMS respondents. Follow-up analyses examined the satisfaction with services between TEAMS and non-TEAMS members. However, no differences were found for any service.

Ordinary least squares (OLS) was used to examine the summary score representing the number of services of which students were aware. After controlling for gender, academic level, and GPA, TEAMS members reported being aware of more services than non-TEAMS members ($R^2 = .048$, $F = 6.57$, $p = .0\ 1$). TEAMS member reported being aware of 7.5 (*SD* = 1.78) services while non-TEAMS members reported being aware of 6.9 (*SD* = 1.92) services.

Satisfaction

Respondents were asked to indicate their agreement with 10 statements concerning the attitude and perception of TEAMS members toward the university and the university community using a 5 point scale (1 = Strongly Agree and 5 = Strongly Disagree). Table 3 shows the statements and the number and percent agreement for TEAMS and non-TEAMS members. Responses of Strongly Agree and Agree were compared to No Opinion, Disagree, and Strongly Disagree for the two groups. Statements 1, 2, 7, and 8 in Table 3 concerned the students' views of the institution. For the first two statements, TEAMS members viewed the university more positively than non-TEAMS members. This positive view held even after controlling for gender, academic level, and student-reported GPA. Compared to non-TEAMS respondents, TEAMS members were 1.6 times more likely to rate new student orientation helpful and 1.3 times more likely to rate the university sensitive to their needs. Although the differences were not statistically significant after corrections were made for multiple tests, in this sample TEAMS members were 1.3 times more likely to agree that the university develops students academically and 1.4 times more likely to agree that they are receiving an adequate education than non-TEAMS respondents. Differences also were examined using OLS regression on a summary score computed as the sum of these four items. The set of predictors accounted for 3% of the variance of the summary score, and TEAMS membership was a significant predictor ($F = 7.65$, $p = .006$).

Statements 4, 5, and 6 in Table 3 all concern the classroom experience and interactions with faculty. No differences were found between the TEAMS and non-TEAMS members for these statements, however, the majority of both groups were in agreement with these statements indicating a positive view of minority students toward their experience with faculty. This positive view was held by minority students regardless of gender, academic level, and student-reported GPA. Differences were not found between TEAMS and non-TEAMS members on a summary score computed from these three items when examined with OLS regression.

Statements 3 and 9 in Table 3 concern activities on campus available for students. No differences were found between groups concerning the number of activities (statement 3). TEAMS members indicated that they feel more included in social activities than non-TEAMS members, however, even then the majority (62.50%) did not feel included. After controlling for gender, academic level, and student-reported GPA, TEAMS members were 1.5 times more likely to agree with statement 9 in Table 3 than non-TEAMS respondents. When the summary score of these two items was examined using OLS regression, TEAMS membership was not a significant predictor.

No difference was found between the two groups for statement 10 in Table 3 in the bivariate (chi-square) analysis, indicating that neither group felt that student government represented their interests. However, after controlling for gender, academic level, and student-reported GPA, TEAMS members were 1.4 times more likely to agree that student government represented their interests than non-TEAMS respondents.

Table 3
Number and Percent of Minority Students that Agree[a] with Attitude Statements Concerning University Experience

Statements	TEAMS member		Non TEAMS member	
	N	*Percent*	*N*	*Percent*
1. New Student Orientation was helpful.	34	47.22*	98	26.42
2. The university is sensitive to student's needs.	36	49.32*	116	31.27
3. There are enough activities ... on campus.	29	19.86	117	31.45
4. Professors are available during office hours.	55	73.33	263	70.51
5. I feel comfortable talking to my professor.	57	77.03	283	75.87
6. I'm able to ask questions during class.	66	88.00	308	82.57
7. Institution develops a student academically/socially.	39	53.42+	148	39.89
8. I feel I'm receiving an adequate education.	65	86.67+	274	73.66
9. I feel included in social activities on campus.	27	37.50*	68	18.38
10. Student government represents my interests.	12	16.67	36	9.76

+ significant before correction for number of tests
* significant after correction for number of tests
[a] Agree responses: 1 = Strongly Agree 2 = Agree

Retention

Enrollment of African-American students increased during the three years after the start of the TEAMS program for both graduate and undergraduate programs at the institution. At a time when total university enrollment dropped by 11 percent, enrollment for African-American students increased 5% for both undergraduate and graduate programs. Retention rates for minority students also improved during this period. For example, retention rates for minority full-time, first-time entering freshmen increased from 63% to 67% between 1993 and 1994. This increase put retention rates for minority students slightly higher than the 66% retention rate for White full-time, first-time entering freshmen.

Although retention rates are not available for all students by classification and race, Table 4 shows the retention rates for first-time, degree seeking freshmen. As Table 1 shows, 63% of the 164 minority students enrolled in 1991 were retained in 1992, the year the TEAMS program began. This apparent gain (over the retention rate of 51% the year before) was not maintained the next year when the retention rate dropped to 55%. Although a general pattern is difficult to establish with only three years of data, for this select group of students, retention rates for minorities appear to be approaching the rate for White students. For the three years prior to the start of the TEAMS program, the average retention rate was 55%, and it was 58% for the three years after the program began.

Table 4
Number and Percent of Students Enrolled and Retained by Minority Status for First Time, Degree-Seeking Freshmen

Year	White Students Head Count	White Students Percent Retained	Minority Students Head Count	Minority Students Percent Retained
1988	816	63%	77	52%
1989	840	62%	152	61%
1990	792	61%	117	51%
1991	738	60%	164	63%
1992	794	61%	240	55%
1993	523	61%	275	56%
1994	462	56%	184	64%
1995	504	64%	215	61%

Of the 171 TEAMS students (including graduate and undergraduate students) tracked over 4 semesters, only 10 (6%) did not enroll in school after the first semester. Eighty-nine percent completed the 94-95 school year, and 71% were enrolled all four semesters or graduated.

Discussion

This study examined the effectiveness of the Teaching Enhancements Affecting Minority Students Program. The program's goal was to increase the retention of minority students by improving the interaction of minority students with other students, instructors, and the university administration. Although there is sufficient literature to support the use of a mentor-based intervention to improve retention rates, its use in nontraditional educational settings is not as clear. The results of this study support the use of the program to improve students' attitudes toward the university community. The program seems less effective in helping students access student services. Although TEAMS students reported being aware of more services than non-TEAMS members, it appears that this difference may be due to only one service. In addition, preliminary analysis indicates that TEAMS may be improving retention rates. The conclusions of this study are consistent with the results of traditional educational settings and the only study available in a nontraditional setting (Lewis, 1986). The African American Freshman Network (BFN) at Georgia State University has also shown positive results from their planned mentoring program.

In order to evaluate the quality of this evaluation, it was examined in light of the program evaluation standards compiled by the Joint Committee on Standards for Educational Evaluation (1994). The Standards address four attributes of an evaluation: utility, feasibility, propriety, and accuracy. The Standards are useful at every stage of evaluation, from implementation through assessing evaluation reports. The utility standards provide guidance to evaluators in efforts to make the evaluation informative, timely, and influential. This evaluation was conducted by an independent evaluator at the request of the administrator responsible for the TEAMS program. Although internal evaluation of the program had been helpful in assessing student satisfaction and needs of students, evidence of program impact was sought to support request for future funding of the program. The feasibility standards address the planning and implementation that may be impacted by cost, time, or political realities. This evaluation was limited to a very narrow scope, because funds were limited. The evaluation attempted to obtain the most useful information with the limited resources available. At the same time, given the scarcity of resources it was important to have some independent information concerning the effectiveness of the program so that funds could be allocated wisely. The propriety standards guided the development of the evaluation plan to assure that students' and program staff's rights were protected legally and ethically. In accordance with these standards, students were informed of the evaluation in a cover letter with the survey, and all responses were confidential.

The accuracy standards are most useful in reporting an evaluation study. These standards are intended to "ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated" (Joint Committee on Standards for Educational Evaluation, 1994, p. 125). Areas in which this evaluation was unable to meet these guidelines are considered limitations. Because the study is based on the volunteer response of students, bias may exist in the responses. However, a comparison of characteristics of the respondents to all African-American students and to non-responding TEAMS members suggest that these differences may not be serious. A more serious concern is the self-selection of students into the TEAMS program. It is not clear how this may impact the results of this study. Students may become involved with TEAMS as a way to effect change in the university system. That is, they may be the more dissatisfied African-American students. Conversely, students joining TEAMS may be the more motivated African-American students and less likely to be dissatisfied. A randomized design in which students were randomly assigned to TEAMS would have addressed this problem, however, this design would run counter to standards addressed in the feasibility standards.

Another concern is that the results of this survey are not a result of the TEAMS program but of some other effort to impact minority attitudes and retention. In 1991, a federally funded program targeted to low-income and first generation students was established. The goal of this program was to help Junior and Senior level students prepare for graduate programs. About 25 students are active in the summer research internship–the main activity of the program–each year. Although this program may also have had an impact on minority retention, attitudes, and knowledge, because it serves a much smaller number of students, has different goals, and targets a different population, the impact would be expected to be small. Another university effort focused on tuition and scholarship money for minority graduate students. Because the tuition and scholarship program focused primarily on graduate students, it is not likely the results of this evaluation were seriously impacted by it.

This study used the traditional assessment format in the study of satisfaction questions (Table 3), and it is possible that the results are not a reliable measure of satisfaction. Franklin and Shemwell (1995) found a clear disparity between this traditional approach and an approach that compared expectations with performance. However, there is no reason to believe that any bias that may exist in the data collection method would not be similar for TEAMS members and non-TEAMS members.

TEAMS has earned the support of the university community and acceptance by the African-American students. TEAMS administration staff have also responded to a continued internal evaluation that has resulted in program changes since students were surveyed in 1994. For example, program staff and students supported a change in the structure of the program. Previously, students were divided by year of education (freshmen and sophomores were separated from upperclassmen). Students are now combined regardless of educational level. These changes require a continued evaluation of the program. During the fall of 1996, the TEAMS program had 221 active members and had further expanded its services to include such things as book grant awards and a literary group for TEAMS members interested in writing. In 1995, the TEAMS program was honored as one of two runners-up for the Council of Graduate Schools/Petersons Award for outstanding programs for recruitment and retention of minority students.

References

Adams, H. G., (1992). *Mentoring: An essential factor in the doctoral process for minority students*. Notre Dame, IN: GEM.

Adams, H. G., (1993). *Focusing on the campus milieu*. Notre Dame, IN: GEM.

Franklin, K. K., & Shemwell, D. W. (1995, November). *Disconfirmation theory: An approach to student satisfaction assessment in higher education*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, Mississippi.

Jacobi, M. (1991). Mentoring and undergraduate academic success: A literature Review. *Review of Educational Research, 61*, 505-532.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards*. (2nd ed.). Thousand Oaks, CA: SAGE, Inc.

Kobrak, P. (1992). Black student retention in predominantly white regional universities: The politics of faculty involvement. *Journal of Negro Education, 61*, 509-530.

Lewis, J. J. (1986). The black freshman network. *College and University, 61*, 135-140.

Mingle J. R. (1987). *Focus on minorities: Trends in higher education participation and success*. Education Commission of the States, Denver, Colo.; State Higher Education Executive Officers Association. (ERIC Document Reproduction Service No. ED 287 404).

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass Publishers.

Peterson, S. (1993). Career decision-making self-efficacy and institutional integration of underprepared college students. *Research in Higher Education, 34*, 659-685.

Redmond, S. P. (1990). Mentoring and cultural diversity in academic settings. *American Behavioral Scientist, 34*, 188-200.

Simmons, R. O., & Thurgood, D. H. (1995). *Summary report 1994: Doctorate recipients from United States universities*. Washington, D. C.: National Academy Press.

Stewart, D. M. (1992). Higher education. In D. W. Hornbeck & L. M. Salamon (Eds), *Human capital and America's future* (pp. 193-219). Baltimore: Johns Hopkins University Press.

Terenzizi, P. T., Rendon, L. I., Upcraft, M. L., Millar, S. B., Allison, K. W., Gregg, P. L., & Jalomo, R. (1994). The transition to college: Diverse students, diverse stories. *Research in Higher Education, 35*, 57-73.

Wiley III, E. (1989). Mentor programs successful in minority retention. *Black Issues in Higher Education, 5*, 8.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 17-23

School Counselors' Perceptions of the Counseling Needs of Biracial Children in an Urban Educational Setting

Nancy J. Nishimura and Linda Bol
The University of Memphis

This study addressed school counselors' perceptions of biracial students (children whose biological parents are of dissimilar racial groups) and the counseling services available to this student population within the school system. Specifically, the study examined the perceptions school counselors hold regarding the counseling needs of biracial children as well as their attitudes regarding whether schools should provide these counseling services. School counselors were surveyed about what they were currently doing in their school to address the counseling needs of biracial children. Survey results suggest that school counselors are satisfied with counseling services currently available to biracial students within the school setting. This study presents another perspective to much of the current professional counseling literature's emphasis on the need for more counseling services.

There has been a recent upsurge of interest in addressing multicultural and/or diversity issues in the mental health literature (Herring, 1992). However, little attention has been directed toward the counseling needs of biracial children (Adler, 1987; Brandell, 1988; Gibbs, 1987; Johnson, 1992a; Kerwin, Ponterotto, Jackson, & Harris, 1993; Nishimura, 1995; Winn & Priest, 1993). In addition, empirical evidence on the racial identity of biracial people is near nonexistent (Tizard & Phoenix, 1995).

The biracial baby boom in this country started in 1967 which was when the last laws prohibiting mixed race marriages were repealed (Root, 1992). Estimates of the number of biracial children in the United States range anywhere from 600,000 to five million (Herring, 1992). This wide-range in estimates is due, in part, to inadequate demographic data collection and, also, a desire for privacy by individual parents (Kerwin & Ponterotto, 1995). Moreover, the number of biracial children is expected to increase (Gibbs, 1987). As public schools continue to experience a similar increase in their biracial student population, school personnel will have increased opportunity for daily interaction with biracial children.

The first reaction of many people when hearing the term *biracial children* is to think of children whose parents are White and African American. In truth, White/African American children are a minority within the biracial children population (Gibbs & Hines, 1992). Most biracial children in the United States are from families where both parents are from two different non-White racial groups. A biracial child is defined herein as one whose biological parents are of dissimilar racial groups, for example, African-American/Asian, Puerto Rican/Native American, White/African-American (Kerwin & Ponterotto, 1995; Winn & Priest, 1993).

The Counseling Needs of Biracial Children

Biracial children struggle with challenges associated with growing up similar to those of their monoracial peers. Erikson (1968) outlined the developmental process in which children in the United States move toward establishing a sense of personal identity. Various foci are highlighted during different developmental stages: self in relation to self, self in relation to family, and self in relation to society. In turn, each stage is examined and attitudes created in response to a newly formed perspective of oneself.

The unique additional challenge for biracial children is that their racial heritage is a combination of two or more racial groups which is often notable in their physical features (e.g., skin color, hair texture, etc.). In a country in which race influences how an individual is perceived by others, what community an individual identifies with, and to some extent, an individual's social relationships, a multi-racial heritage has the potential to present a complex situation (Phinney, 1989; Spickard, 1992; Tizard & Phoenix, 1995). For monoracial children, racial self-labeling usually coincides with racial labeling by others. Devoid of racial identity incongruity, the children are then free to move to the next stage of processing what that

---

Nancy Nishimura is Assistant Professor in the Counseling, Educational Psychology, and Research Department at The Memphis State University. Linda Bol is Assistant Professor in the same department. Please direct correspondence regarding this article to Nancy J. Nishimura, 100 Ball Hall, The University of Memphis, Memphis, TN 38152 or by e-mail at nishimura.nancy@coe.memphis.edu racial label means in terms of the person within themselves.

Biracial children often struggle with an additional developmental step in which racial identity must be examined. Many biracial children experience pressure to make a decision to declare a specific racial identity (Bradshaw, 1992; Kerwin & Ponterotto, 1995; Tizard & Phoenix, 1995). One option that was popular during the 1980's was for biracial children who have an African American parent to assume a Black identity to develop a healthy sense of self (Tizard & Phoenix, 1995). Other options that have been promoted by various social groups and institutions are that biracial children's racial identity is determined by: (a) the racial identity of the father, or (b) the racial identity of the mother. As a result of these inconsistencies, biracial children often operate under different racial identities depending on the social context. Identity selection may occur with a certain degree of discomfort, as the children often feel that one parent is being negated, as a result of having to claim one racial heritage over the other (Kich, 1992).

For example, a child of African American/Jewish European American parentage may socialize and identify closely with African American classmates at school. At home, this same child may identify closely with her Jewish parent, carefully following Jewish traditions in family interactions. In the African American or the Jewish community at large, this child may not be viewed as a full member, neither Black nor Jewish enough to suit certain community members (Johnson, 1992b). Biracial children are compelled to sort through who they are while struggling to process societal messages about what they are. This dilemma simultaneously is, at best, a difficult process for children to negotiate.

Although there is a good deal of literature to support the argument that biracial children have special counseling needs, no empirical data are available to support this claim (Gibbs, 1987; Herring, 1992; Kerwin & Ponterotto, 1995; Kerwin, Ponterotto, Jackson, & Harris, 1993; McRoy & Freeman, 1986; Nishimura, 1995).

The Role of the School Counselor

School counselors have been urged to address the counseling needs of biracial children (Herring, 1992; Nishimura, 1995). Specific strategies include co-facilitating classroom activities with teachers, conducting individual counseling sessions, facilitating support groups for students, and initiating parent networking groups (Wardle, 1992). The benefits of the aforementioned strategies include: (a) addressing identity development (including racial identity) for a growing segment of the student population, which is already a priority in developmental school counseling programs (Department of Education/Indiana School Counseling Association, 1991; McRoy & Freeman, 1986); (b) promoting sensitivity and appreciation of diversity for all students (Wardle, 1992); and (c) fostering stronger ties with parents and the community (Cole, Thomas, & Lee, 1988).

A two-pronged focus is evolving in the literature which emphasizes: (a) the need to address the counseling issues of biracial children and (b) the encouragement of school counselors to take on that challenge (Kerwin & Ponterotto, 1995). In reviewing the literature, what seems to be missing is an assessment of school counselors' perceptions about the counseling needs of biracial children and how they interpret their role in addressing these needs. Therefore, the need for counseling focus has been articulated in the literature, and the service providers (school counselors) have been identified and encouraged to take action. However, without data on whether and how counseling services are provided to biracial children in the school setting, the implementation phase remains an unknown.

The research questions addressed in this study are as follows: (1) What perceptions do school counselors hold regarding the counseling needs of biracial children? (2) What are school counselors currently doing in their school to address the counseling needs of biracial children? and (3) Do school counselors believe that their school should provide counseling services highlighting biracial children's counseling needs?

Method

Participants

A questionnaire was mailed to each of the 238 elementary, middle school, and high school counselors employed in an urban school district located in the mid-southern area of the United States. Student enrollment in this district is 108,590 students. The racial composition of all students in the district is 81% African American, 18% European American, and 1% "Other." Of the 238 counselors surveyed, 120 (50%) returned a completed questionnaire.

Measure

The questionnaire was developed by the researchers for the purposes of this study. Item development was guided by the literature on issues of counseling biracial children.

The 13-item questionnaire consisted of four major sections. To ensure a common definition of biracial students, the definition used for the purposes of the survey appeared at the top of the first page of the questionnaire. A biracial student was defined as "one whose biological parents are of dissimilar racial groups (for example, African American/Asian, Puerto Rican/Native American, White/African American)." The first section contained six demographic items including gender, age, ethnicity, years employed as a school counselor, grade levels served by their schools (elementary, middle, or high school), and the presence of biracial students in their assigned schools. The second section asked counselors to describe any difficulties or counseling issues faced by biracial students in comparison to other students in the school. These were four multiple-choice type items with three response options (less, same, or more, for Items 7-9; and unique issues, common issues, or both for Item 10). The third section contained three items on whether: (a) there is a need to incorporate biracial issues in counseling/instructional programs; (b) biracial issues are currently incorporated into their school's counseling/instructional programs, and; (c) the counselor considers him or herself able to meet the counseling needs of biracial students in the existing school counseling program. The response options for these last three items were simply "yes" or "no."

Procedure

After obtaining a roster of all school counselors in the school district from the central administration office, the questionnaire, a cover letter, and a self-addressed stamped envelope were mailed to all school counselors. In the cover letter, the counselors were informed about the purpose of the survey and assured that their responses would remain confidential. The counselors were asked to complete and return the questionnaire within two weeks. After the two weeks had elapsed, follow-up calls were made to each of the counselors reminding them to complete and send the survey if they had not already done so. During the phone conversations, the researchers offered to send another copy of the questionnaire to those counselors who had misplaced the original instrument.

Analysis

In addition to obtaining descriptive data for each item, chi square analyses were computed to explore the relationships between demographic variables and responses to items about the counseling issues and the counseling needs of biracial students. Chi square analyses were also used to explore the relationships between counseling issues and the counseling needs of biracial students. Chi squares were used to analyze the data because responses on all variables were categorical in nature.

Results

Demographic Characteristics of Respondents

The respondents were predominately women; the sample consisted of 106 (88%) women and only 11 (9%) men. Three persons did not identify their gender. Over half of the respondents were African American (57%), followed by European American (34%), and Native American (6%). One person identified herself as biracial, while three persons did not provide ethnicity information. In reference to the age groups of respondents, the largest percentage (44%) was between 46 to 53 years of age. Twenty-seven percent of respondents fell between the ages of 54 to 69, 18% were between 38 to 45, and 10% were 37 years or younger. As for school assignment, 51% of counselors were employed at elementary schools, 18% were at middle schools, and 30% were assigned to high schools. The largest percentage of respondents had less than five years of experience working as a school counselor (31%), while the second largest percentage had over 20 years of experience (24%). The remaining percentages by years of experience were similar, with 18% of school counselors having 6-10 years, 14% having 11-15 years, and 13% having 16 to 20 years of experience.

Survey respondents were representative of counselors employed by the school district. According to the coordinator of the school district's secondary school counseling program, of the 238 school counselors employed by the district, there are 212 (89%) women and 26 (11%) men. Over half of the school counselors are African American (57%), followed by European American (42%), and Asian (0.4%). As for school assignment, 51% of the counselors are employed at elementary schools, and 49% are assigned to secondary schools (middle school and high school). Though there were no data collected for the other demographic variables, the district percentages obtained for gender, ethnicity, and school assignment are nearly identical to the demographic information obtained in the present sample.

Nearly all respondents indicated that there were biracial students enrolled in their school. Ninety-three percent ($n = 111$) of the counselors responded that there were biracial students in their school, while only 7% ($n = 9$) indicated that they did not have biracial students on campus.

Counseling Issues of Biracial Students

The counseling issues of biracial students were addressed by surveying school counselors' perceptions regarding the extent of problems these students experience in the school setting. The descriptive statistics for the first three of these items appear in Table 1.

The results for the first item indicate that most school counselors think that biracial students experience the same behavioral problems experienced by other students (78%). Only 17% of the counselors responded that biracial students experience more behavioral problems. A similar pattern was obtained for the item on whether biracial students experience more problem issues when compared to other students. Seventy-three percent of the sample said they experience the same kinds of problems in counseling, and only 20% indicated that biracial students experience more problems when compared to their other students. A large percentage of counselors perceived that biracial students had more difficulty with peer acceptance than other students (40%). However, the largest percentage of respondents still perceived that biracial students had the same amount of difficulty in the area of peer acceptance. On the final item in this section, counselors were asked to indicate whether the counseling issues faced by biracial students were unique, common, or both. Fifty-seven percent of respondents considered the counseling issues of biracial students to be both unique and common, while 31% considered them to be common to all students. Only 12% of the counselors thought that biracial students experienced unique counseling issues.

Table 1
Frequencies and Percentages of Respondents by Response Category on Items Related to the Counseling Issues of Biracial Students

Item	Less problems/ difficulty n	Less problems/ difficulty %	Same problems/ difficulty n	Same problems/ difficulty %	More problems/ difficulty n	More problems/ difficulty %
7. Behavioral problems experienced by biracial students in comparison to other students.	6	5	93	78	21	17
8. Peer acceptance difficulties experienced by biracial students in comparison to other students.	8	7	63	53	48	40
9. Problems in the presentation of counseling issues experienced by biracial students compared to other students.	8	7	87	73	20	20

Among the questions addressed by the chi square analyses was whether the respondents would judge the extent of problems or difficulties faced by biracial students differently depending on their own backgrounds. That is, would the ethnicity, age, years of experience, and school placement of the counselors be related to their perceptions of the extent of problems experienced by biracial students when compared to other students? There were too few male counselors in the sample to allow a valid comparison by gender. The results revealed one significant difference in responses to these items based on any of these demographic variables. There was a significant difference in the perceived extent of behavioral problems (Item 7) depending on school assignment [$\chi^2(4, N=116)=16.44, p=.002$]. School counselors assigned to the middle schools were more likely to report that biracial students experience more behavioral problems (See Table 2). Apparently, the other background characteristics of the school counselors were not related to their perceptions of the counseling issues experienced by biracial students compared to other students.

Counseling Needs of Biracial Students

The items addressing the counseling needs of biracial students asked respondents to indicate whether: (a) there is a need for instructional/counseling programs with an emphasis on biracial issues, (b) such a program is currently implemented in their schools, and (c) the school counselor is able to meet the counseling needs of biracial students in their present school counseling programs. The frequencies and percentages of school counselors who responded affirmatively or negatively to these items appear in Table 3.

In response to the question about whether there is a need to incorporate biracial issues into instructional/counseling programs, most respondents replied that there was no need for this type of program in their schools.

Whereas 76% of school counselors replied there was no need for this type of program, only 24% agreed that this need existed. Not only did counselors judge there was no need for such a program, an even larger percentage (84%) reported that their schools are not presently implementing counseling/instructional programs that address biracial issues. Sixteen percent of school counselors reported that such a program exists in their school. On the final item in this section, 90% of respondents judged that they were able to meet the counseling needs of biracial students within the present counseling program without modification; only 10% disagreed with this statement.

Table 2
Frequencies and Percentages by Response Category for Significant Chi Square Results

		Item 7: Extent of behavioral problems experienced by biracial students			
		Less	Same	More	Total
School Assignment	Elementary	5 (8%)	44 (73%)	11 (18%)	60
	Middle	0 (0%)	11 (58%)	8 (42%)	19
	High	1 (3%)	35 (95%)	1 (3%)	37

		Item 11: Need for biracial counseling program		
		Yes	No	Total
Item 9: Extent of problems experienced by biracial students in counseling	Less	0 (0%)	8 (100%)	8
	Same	16 (19%)	67 (81%)	83
	More	10 (42%)	14 (58%)	24

		Item 11: Need for biracial counseling program		
		Yes	No	Total
Item 7: Extent of behavioral problems experienced by biracial students	Less	1 (17%)	5 (83%)	6
	Same	16 (18%)	73 (82%)	89
	More	9 (43%)	12 (57%)	21

Table 3
Frequencies and Percentages of Respondents by Response Category
on Items Addressing the Counseling Needs of Biracial Students

Item		Yes n	Yes %	No n	No %
11.	In my school there is a need to incorporate into instructional/counseling programs an emphasis on various issues related to being biracial.	28	24	90	76
12.	My school is presently implementing counseling/instructional programs which address various issues related to being biracial.	19	16	101	84
13.	I am able to meet the counseling needs of biracial students within the present counseling program in my school without modification.	106	90	12	10

The chi square analyses used to investigate whether the school counselors' responses to the items about the counseling needs of biracial students varied as a function of the demographic variables did not yield any significant findings. The respondents' ethnicity, age, years of experience, or school assignment did not predict their perceptions of the need for incorporating biracial issues in school counseling programs.

However, significant results were obtained when responses to items about the counseling issues of biracial students were compared to the responses on items about the counseling needs of biracial students. In other words, there seemed to be some meaningful relationships between how school counselors perceived the problems or difficulties experienced by biracial students and their judgment of the counseling needs of these students.

The first significant finding was a comparison of whether the respondents saw a need for a biracial counseling program and their perception of the extent of problems faced by biracial students in counseling [$\chi^2(2,N=115)=7.84, p=.02$]. The number and percentage of respondents by response category are provided in Table 2. An examination of this table reveals that respondents are more likely to say there is a need for a biracial counseling program if they judge that biracial students experience more problems in counseling than do other students.

A similar pattern of results was found when comparing responses to Item 7 and to Item 11 (see Table 2). There was a significant result when comparing responses about the need for biracial student counseling and perceptions about the extent of behavioral problems experienced by biracial students [$\chi^2(2,N=116)=6.17$, $p=.046$]. The respondents were more likely to endorse the need for a biracial counseling program if they judged that biracial students experienced more behavioral problems than other students.

Discussion

Results of the study highlighted previously untapped school counselor perceptions regarding the counseling needs of biracial students and their role as service provider to this population. Though school counselors did not perceive that biracial students have more counseling related problems than their monoracial peers, there was some recognition that the area of peer acceptance posed significant challenges for biracial students. However, if developing harmonious peer relationships is a large part of the developmental process of most young people in this society, then perhaps the challenge it poses for biracial students can be more readily appreciated (Erikson, 1968).

School counselors also indicated that biracial students presented both general and unique counseling issues. This information supports the premise that biracial students are faced with coping with the normal challenges of growing up as well having to struggle with issues unique to their situation.

There was a strong opinion conveyed in the survey that while their school was not promoting any instructional/counseling program that emphasized issues related to being biracial, school counselors, as a group, did not support the need to implement such programs. The respondents clearly indicated that they were able to meet the counseling needs of biracial students within their present counseling program without modification.

Results of this study suggest a difference in perceptions between service providers (school counselors) and the literature regarding the counseling needs of biracial students (Gibbs, 1987; Herring, 1992; Nishimura, 1995; Wardle, 1992). This exploratory study suggests that there may exist different perspectives regarding the counseling needs of biracial children, even among groups which hold as a priority serving the needs of the whole child in order to facilitate development. Without a mutual sharing of perspectives and feedback, various entities (e.g., service providers and the professional literature) may end up operating in a vacuum. If school counselors are satisfied regarding the counseling services they provide biracial students and do not make changes in service delivery, it is conceivable that biracial students will not receive situation specific counseling services in the schools. On the other hand, the professional literature may continue to call for counseling services when they are unnecessary. While resolution of this dilemma awaits empirical confirmation, it is suggested that communication lines between school counselors and counselor educators/theorists remain open and facilitative on this issue.

Some limitations of the present study should be noted. The first is whether the counselors who responded are representative of all counselors in the school district. Though the demographic characteristics of respondents were similar to the demographic characteristics of non-respondents, the possibility of sample bias cannot be overruled. Similarly, the results obtained from this sample of counselors may not be generalizable to counselors working in different cities and in different types of schools. It may be that counselors in schools with different proportions of students from various ethnic groups would perceive the counseling issues and needs of biracial students very differently. A final limitation was the exploratory nature of the data analysis. However, the relationships we observed were sensible and easily interpretable, even though these interpretations were post-hoc. The authors suggest that state, regional, and national surveys of school counselors be conducted using a lengthier questionnaire to provide perspective for the preliminary findings of this study. The end result of these described dynamics is that the students, themselves, may be the ultimate loser by missing the opportunity to benefit from an increase in support, social awareness, and appreciation and sensitivity of multicultural aspects of themselves and others.

References

Adler, A. J. (1987). Children and biracial identity. In A. Thomas & J. Grimes (Eds.), *Children's needs: Psychological perspectives* (pp. 556-66). Washington, DC: The National Association of School Psychologists.

Bradshaw, C. K. (1992). Beauty and the beast: On racial ambiguity. In M.P.P. Root (Ed.), *Racially mixed people in America* (pp. 77-88). Newbury Park, CA: Sage Publications.

Brandell, J. R. (1988). Treatment of the biracial child: Theoretical and clinical issues. *Journal of Multicultural Counseling and Development, 16*, 176-187.

Cole, S. M., Thomas, A. R., & Lee, C. C. (1988). School counselor and school psychologist: Partners in minority family outreach. *Journal of Multicultural Counseling and Development, 16*, 110-116.

Department of Education and Indiana School Counselors Association. (1991). *A model for developmental school counseling programs in Indiana*. Indianapolis, IN: Author.

Erikson, E. (1968). *Identity: Youth and crisis*. New York: Norton.

Gibbs, J. T. (1987). Identity and marginality: Issues in the treatment of biracial adolescents. *American Journal of Orthopsychiatry, 57*, 265-278.

Gibbs, J. T. & Hines, A. M. (1992). Negotiating ethnic identity: Issues for Black-White biracial adolescents. In M.P.P. Root (Ed.), *Racially mixed people in America* (223-238). Newbury Park, CA: Sage Publications.

Herring, R. D. (1992). Biracial children: An increasing concern for elementary and middle school counselors. *Elementary School Guidance and Counseling, 27*, 123-130.

Johnson, D. J. (1992a). Racial preference and biculturality in biracial preschoolers. *Merrill-Palmer Quarterly, 38*, 233-244.

Johnson, D. J. (1992b). Developmental pathways: Toward an ecological theoretical formulation of race identity in Black-White biracial children. In M. P. P. Root (Ed.), *Racially mixed people in America* (pp. 37-49). Newbury Park, CA: Sage Publications.

Kerwin, C. & Ponterotto, J. G. (1995). Biracial identity development. In J. G. Ponterotto et al. (Eds.), *Handbook of multicultural counseling* (pp. 199-217). Thousand Oaks, CA: Sage Publications.

Kerwin, C., Ponterotto, J. G., Jackson, B. L., & Harris, A. (1993). Racial identity in biracial children: A qualitative investigation. *Journal of Counseling Psychology, 40*, 221-231.

Kich, G. K. (1992). The developmental process of asserting a biracial, bicultural identity. In M.P.P Root (Ed.), *Racially mixed people in America* (pp. 304-317). Newbury Park, CA: Sage Publications.

McRoy, R. G. & Freeman, E. (1986). Racial-identity issues among mixed-race children. *Social Work in Education, 8*, 164-174.

Nishimura, N. (1995). Addressing the needs of biracial children: An issue for school counselors in a multicultural school environment. *School Counselor, 43*, 52-57.

Phinney, J. S. (1989). Stages of ethnic identity development in minority group adolescents. *Journal of Early Adolescence, 9*, 34-49.

Root, M. P. P. (1992). Within, between, and beyond race. In M. P. P. Root (Ed.), *Racially mixed people in America* (pp. 3-11). Newbury Park, CA: Sage Publications.

Spickard, P. R. (1992). The illogic of American racial categories. In M. P. P. Root (Ed.), *Racially mixed people in America* (pp. 12-23). Newbury Park, CA: Sage Publications.

Tizard, B. & Phoenix, A. (1995). The identity of mixed parentage adolescents. *Journal of Child Psychology and Psychiatry, 36*(8), 1399-1410.

Wardle, F. (1992). Supporting biracial children in the school setting. *Education and Treatment of Children, 15*, 163-172.

Winn, N. & Priest, R. (1993). Counseling biracial children: A forgotten component of multicultural counseling. *Family Therapy, 20*(1), 29-36.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 25-38

How Experienced Teachers Think About Their Teaching: Their Focus, Beliefs, and Types of Reflection

Rita M. Bean, Deborah Fulmer, Naomi Zigmond
University of Pittsburgh

Judith V. Grumet
Gateway School District

The purpose of this study was to describe how four experienced elementary teachers reflected on a lesson that they had taught and then viewed on videotape. We investigated (1) how teachers used their knowledge of social studies, pedagogy, and experience in thinking about their lessons, and (2) the emphases they placed on content, instruction, students, and classroom management. Teachers most frequently used both experience and theory to explain their instructional approaches. They had strong beliefs about instruction. These teachers identified strengths of their lessons, erroneous assumptions, and changes that they would make if they were to teach the lesson again. Teachers appreciate the opportunity to participate in the reflective process as a means of improving their own teaching.

Recently there has been a great deal of interest in having teachers reflect upon their own instruction as a means of helping them further their own understandings and think more deeply about their profession (Denton & Peters, 1988; Russell, 1993; Schon, 1983). Reflection is not new however; it derives from the ancient idea that wisdom is comprised of the ability to analyze situations, recognize nuances posed by problems, think diligently, and propose solutions (Houston, 1988). As Socrates noted "Knowledge is sought within the mind and is brought to birth by questioning" (Knapp, 1992).

There have been several attempts to define reflection, with Schon's work a touchstone in the literature (1983). In fact, Schon's two types of reflection, Reflection-on-Action (reflection on practices, action, and thoughts after the practice is completed) and Reflection-in-Action (reflection on phenomena and on one's spontaneous ways of thinking and acting in the midst of action) have given rise to a third type of reflection, Reflection-for-Action, (reflection in order to guide future action) as suggested by Killion and Todnem (1991). The Killion and Todnem model evolved from workshops in which they invited teachers to describe their own work, develop understanding of patterns in their own behaviors, establish cause-effect relationships between their actions and the outcomes observed, and construct a rationale for their work.

Based in part on the writings of Zeichner and Liston (1987) and Van Manen (1997), Sparks-Langer, Simmons, Pasch, Colton, and Starko (1992) developed a Framework for Reflective Pedagogical Thinking. This framework was used in a pre-student teaching program to evaluate students' ability to reflect on the various components of the curriculum (Sparks-Langer, Simmons, Pasch, Colton & Starko, 1992). In discussing this implementation, Sparks-Langer et al. (1992) acknowledged difficulties with the fact that their framework implied a stable and linear growth in reflection. They suggested that a dual coding system was needed, one code for technical thinking and another code for moral/ethical thinking. Ross (1988) discussed another theoretical framework for defining reflection which includes a progressive development of competence in making reflective judgments. Ross argues that this simple framework is helpful in describing qualitative changes in the progress of reflective judgment including: (1) development of the processes involved in reflection, (2) development of attitudes essential to reflection, and (3) development of the appropriate content of reflection. Ross found that levels of reflective judgment increased with both age and education, with the highest levels seen in advanced graduate students.

Reflection, then, is based on knowledge and professional experience (Canning, 1991; Hayes & Ross, 1988; Moore, Mintz, & Biermann, 1988; Munby & Russell, 1994; Van Manen, 1977; Zeichner, 1990). In fact, a knowledge base that includes pedagogical theory is a prerequisite to reflective practice (Adler, 1994; Griffin,

---

This research was supported, in part, by a grant (H#023D00003, Analysis of Social Studies Curriculum and Instruction for Mainstream and Learning Disabled Students) from the Office of Special Education Programs, Department of Education. Rita M. Bean is Professor and Associate Dean, Deborah Fulmer is Research Specialist, and Naomi Zigmond is Professor, School of Education, University of Pittsburgh. Judith V. Grumet is a teacher with the Gateway School District (Monroeville, PA). Correspondence regarding this article should be directed to Rita M. Bean, University of Pittsburgh, 5T23 Forbes Quadrangle, Pittsburgh, PA 15260 or by e-mail to bean@fs1.sched.pitt.edu.

1991; Russell, 1993; Smith, 1991; Zeichner, 1990; Zeichner & Liston, 1987). In addition, to be reflective, teachers require instruction about what to reflect on and how to be productive and effective in reflection (Askling & Almen, 1995; Denton & Peters, 1988; Evans, 1991; Oja, 1991; Pugach & Johnson, 1988; Richert, 1991). The literature demonstrates that growth in reflection is a constructive process and occurs as a function of peers or teachers and mentors making sense together. Various strategies have been used to promote higher levels of reflection. Peer coaching is thought to help teachers become more reflective (Sparks-Langer et al., 1992). Viewing videotaped lessons and discussing impressions can reveal things of which teachers may have been unaware (Wildman & Niles, 1987). Other approaches to encourage reflective practice include journal writing (Canning, 1991; Holly, 1983; Smith, 1991; Sparks-Langer & Colton, 1991), forum discussions (Evans, 1991), peer teaching, portfolios, peer partnerships, ethnographic studies (Smith, 1991), case studies (Richert, 1991; Smith, 1991), critical dialogue, field experiences (Sparks-Langer & Colton, 1991), action research projects (Smith, 1991; Sparks-Langer & Colton, 1991), and examination of one's belief systems (Marshall, 1990).

The interest in teacher reflection is consistent with the concern for school reform. As Thornton (1994) indicates, teachers tend to be "curricular-instructional gatekeepers" (p. 5). What they do may be based on reflection or upon unexamined assumptions or conventions. And although the different kinds of reflection have been examined extensively in recent years, there appears to be a continuing need to investigate how reflection can contribute to both professional development and improved schooling. To this end, recent research has explored the uses and context of reflective practice.

Social studies educators have been especially interested in reflective practice as a means of promoting teacher growth and professionalism (Ross, 1994). Given the emphasis in social studies on involving learners in decision making and problem-solving about complex issues, past and present, this content field seems particularly suited to an investigation of how experienced teachers think about and reflect on their practices.

In this study, four experienced elementary teachers were given an opportunity to observe a videotape of one lesson they had conducted and to reflect on that lesson in a discussion with the investigators. We were interested in understanding the degree to which these four teachers, who had participated in a year long study of their social studies instruction, relied on their past experiences or integrated these experiences into workshop content that we had provided for them. We were also interested in learning more about the foci of the reflections of these experienced elementary teachers, that is, what emphases they placed on topics such as content, instruction, students, and classroom management, when they viewed and talked about their lesson.

Methods

Teachers

Four elementary teachers participated in this study. The teachers were part of a multi-year comprehensive research project funded by the U.S. Office of Education, Special Education Department, in which mainstream social studies curricula were explored to gain a clearer understanding of the scope, sequence, and presentation of content that produce effective learning in all students who find themselves in mainstream social studies classrooms, including students with learning disabilities. The four teachers provided social studies instruction to students in third, fourth, fifth, and sixth grades. The third and sixth grade teachers taught in the rural district located approximately 40 miles from a large eastern metropolitan city. This district of 2,687 students had approximately 8% of the student body identified as eligible for special education services. The fourth and fifth grade teachers taught in the suburban district, located 10 miles north of the metropolitan area, which served approximately 4,571 students with 6% of students identified as eligible for special education services.

Teaching experience among the four teachers ranged from 17 to 20 years. All four teachers had majored in elementary education. One teacher had a doctorate, another was in a doctoral program, and the remaining two had studied at the graduate level. Two teachers were female, two were male; all were white. These four teachers had volunteered to participate in the study of reflection.

Staff Development Project

As part of this multi-year study, these four teachers (and eight others) attended a 5-day summer workshop in which they participated in activities focused on improving their social studies instruction and helping them gain strategies by which they could adjust instruction for students with learning disabilities who were mainstreamed into their classrooms. Workshop activities were based on a set of guidelines developed by the research team. These guidelines included ideas about planning, teaching, and evaluating social studies lessons. Specific ideas were given about prior knowledge and its importance in learning; use and limitations of social studies textbooks; helping students organize information; active involvement using a lesson framework; vocabulary instruction; grouping to provide for individual differences; and evaluation strategies. As part of this summer workshop, participants chose a unit of study they planned to teach in the Fall semester and modified it to include the ideas they were learning in the workshops. This modified unit was discussed with one of the research teams and suggestions were made for additional changes, as necessary.

During the Fall, teachers were observed by trained project staff, as they taught that modified unit (generally over a 2-week period). Observers took comprehensive narrative field notes describing content, instruction, and student responses. Teachers were also audio-taped using a micro-cassette recorder and lapel microphone to obtain a complete record of teacher and student talk. Various artifacts of instruction were collected including student work samples and teacher lesson plans. During the unit, one lesson was videotaped. These videotapes were the stimulus materials for the interview that captured teacher reflection.

Interview Procedures

Approximately 5 months after the videotaping, volunteers were solicited from the set of 12 teachers to participate in the reflection study. Each of the four teachers who volunteered was sent a set of directions and a copy of the videotape of their lesson. The elapsed time between the videotaped lesson and the viewing was purposeful to allow some distance from the event so that teacher participants were able to reflect on the content of the videotaped lesson of instruction. Teachers were asked to view their video in its entirety, and then think about and respond to a set of questions. Specifically, teachers were asked to identify teaching episodes on the videotape which they characterized as successful, episodes which triggered questions about their teaching practices, episodes in which they demonstrated strategies learned from the project workshop experiences, and any other episodes that they thought were especially interesting. On a scheduled date, each of the four teachers met and discussed their responses with the two project directors. During these individual interviews, project directors guided participants through the videotaped lessons, asking them to reflect on various elements of the lessons. Videotapes were played and replayed, with the teachers identifying various episodes and reflecting on their responses. Transcriptions of these interviews form the data base for this study.

Analysis

The interviews with the four teachers were tape recorded, transcribed, and then converted to *The Ethnograph* (Seidel, 1988) for ease in coding and categorizing. Interview transcripts ranged from 905-1,200 lines of transcription.

For this study, we described reflection as the practice of analyzing one's actions, decisions, or products by focusing on the process of achieving them (Killion & Todnem, 1991). Codes were developed to capture two aspects of reflection, focus and type. There were four categories of *focus*: (INST) Instruction, (CON) Content, (STUD) Students, and (MANAGE) Management (See Figure 1). *Instruction* refers to statements about how teaching was implemented, as well as statements about the materials or grouping strategies used to enhance instruction (i.e., teacher discussion of the use of small groups to promote active involvement would be coded as INST). *Content* refers to statements about what was to be learned, including concepts, vocabulary, and specific details. For example, a teacher reference to the importance of learning how the religious beliefs of the Egyptians influenced their behavior would be coded CON. *Students* refer to any statements about student characteristics or behavior, or adaptations that were made for particular students (i.e., a teacher statement that "Susie's attention span is short and I have to keep involving her" was coded STUD). Finally, *Management* refers to teacher statements about non-instructional concerns such as discipline, interruptions, or student behavior (i.e., a teacher statement that "I was concerned about the group being off-task during one part of the lesson" would be coded MANAGE).

Instruction (INST) - teacher statements about how teaching/learning was implemented (e.g., use of oral reading; use of writing, etc.) May also include materials such as textbooks, grouping strategies

Content (CON) - teacher statements about what was to be learned (concepts, vocabulary, understandings)

Students (STUD) - teacher statements based on attention to student characteristics or needs, adaptations made due to student needs

Management (MANAGE) - teacher reference to non-instructional concerns in the classroom context (discipline, interruptions, student behavior)

Figure 1. Foci of Reflection Statements

In generating a scheme for coding the *types* of reflection teachers engaged in, we made two assumptions. First, experienced teachers tend to use both theoretical principles and their own knowledge from years of classroom teaching; a hierarchical model which set one of these above the other would not be workable. Second, since we were interested in the relationship between the types of reflection and the focus, our analysis scheme would need to allow for coding of both.

A three category coding system was developed (See Figure 2) for identifying three types of reflection: *Description (D)*; *Explanation Based on Experience or Personal Belief (EE)*; and *Explanation Based on Theory or Principle (ETP)*. *Description* refers to a reflection in which the teacher described activities or events, without any elaboration or discussion (e.g., "I used cooperative groupings for today's lesson" was coded as *(D)* to denote a reflection which fell into this category). *Explanation Based on Experience or Personal Belief (EE)* refers to a reflection in which the teacher gave a rationale that seemed to be based on personal beliefs or past experience (e.g., "I always have students read aloud; it helps them understand the textbook"); when a teacher suggested a change in instruction, but did not refer to a particular theory or principle as the reason for the change, we coded the statement as (EE). *Explanations Based on Theory or Principle (ETP)* refer to statements in which a teacher presented a rationale based on some principles or theories related to practice, or some social/ethical/moral dimensions of teaching or learning (e.g., "The class needed more review of material in order to thoroughly understand it."). Ideas for change based on theory were also coded in this category.

1. Description (D) - labeling of events or actions, using appropriate terms, with little explanation or discussion about the events (e.g., used groups, students read orally). Respondent may also raise questions about the lesson (e.g., Did I talk too much?).
2. Explanation based on experience or personal belief operating within the context of the classroom (EE) - elaborates on events or actions and alludes to or presents rationale that is based on experience or personal beliefs germane to the context of the particular classroom (e.g., I did a similar lesson in this class before and it was too long; I think having these students read aloud helps them understand that textbook). Teacher may also suggest change in the lesson as part of this type.
3. Explanation based on theory or principle (ETP) - elaborates on events or actions and presents or alludes to a rationale that is based on some theory or principle related to practice, or some social/ethical/moral dimensions of teaching/learning (e.g., The class needs to review material to learn it).

Figure 2. Codes for Types of Reflections

Coding transcripts was a three-step procedure. First we defined each *episode* as a complete statement made by the teacher following a comment or question by one of the interviewers. Next each episode was coded into one of the three types of reflection. Then the episode was coded for focus. If more than one focus emerged within the episode, the same reflection type was assigned to each focus to preserve a 1:1 ratio between the type and focus of reflections. For example, the following episode was coded as ETP/INSTR (*Explanation Based on Theory or Principle* with a focus on instruction) and ETP/STUD (*Explanation Based on Theory or Principle* with a focus on students).

> I thought that the discussion was very well laid out and beneficial. I thought that . . . I like that kind of a format, where you talk with the kids about the information, read from the text, review the vocabulary. I think you get a very good feel for what they understand when you're doing it with them like that. (Charles)

Three members of the research team coded the four teacher transcripts. Inter-rater reliability was accomplished following the initial coding of documents. One member of the research team re-examined 30% of the coded discussion episodes (i.e., of the 213 episodes identified throughout the 4 transcripts, 64 random discussion episodes were re-coded). Inter-rater agreement was 83%.

To summarize the codes by teacher, we developed a matrix in which we displayed the number of codes by type and focus (See Table 1). For example, 18 of Abby's reflection episodes were coded as *Description (D)*; the focus of these reflections included Instruction (2), Content (7), Students (8), and Management (1).

Results

First, we describe each of the interviews with the teachers and provide a summary of the types and foci of their reflections. This is followed by a summary of themes across all of the teachers.

Abby

For Abby, a third grade teacher, the videotaped lesson occurred midway through a unit of social studies instruction on Map Skills, the 5th day of a 10-day unit. The main objectives for this day's lesson were for students to be able to name and locate the seven continents and the four oceans. Students were directed to engage in a paired reading activity using the text to find meanings of words, determine what countries were located north and south of U.S. borders, and list all seven continents. A follow-up activity involved students labeling a color-coded world map with the continents and oceans, and producing a map key.

An analysis of the post-lesson interview with Abby revealed that the majority of her reflections (37%) were the type referred to as *Explanation Based on Experience or Personal Belief (EE)*. The primary focus of these reflections was on *Students* (13/21) (See Table 1). Abby described the characteristics of various students to explain how she attempted to meet the needs of students at-risk for academic success. The following are episodes of reflection coded EE/STUD.

Table 1
Level of Reflection and Foci of Instruction

Teacher Grade	Types of Reflection	Focus of Instruction INST n %	 CONTENT n %	 STUD n %	 MANAGE n %	Total %
Abby						
3rd	D	2	7	8	1	18 (32)
	EE	6	1	13	1	21 (37)
	ETP	12	3	3	0	18 (32)
		20 (35)	11 (19)	24 (42)	2 (4)	57
Barbara						
4th	D	2	1	2	0	5 (15)
	EE	6	2	9	1	18 (53)
	ETP	5	4	2	0	11 (32)
		13 (38)	7 (21)	13 (38)	1 (3)	34
Charles						
5th	D	5	3	7	0	15 (22)
	EE	7	10	5	0	22 (33)
	ETP	14	8	7	1	30 (45)
		26 (39)	21 (31)	19 (28)	1 (2)	67
David						
6th	D	4	6	3	1	14 (25)
	EE	10	4	2	2	18 (33)
	ETP	13	5	4	1	23 (42)
		27 (49)	15 (27)	9 (16)	4 (7)	55

> . . . And I think there are some kids you need to spend a lot of time with and there are other kids, that, they learn in spite of the teacher. You know, I don't care what you do, the knowledge is there and they've got the skills and abilities to do it. And just keeping them challenged and motivated enough to continue. (EE/STUD)

> She's trained me. You're constantly focusing back on her, checking to make sure she's doing what she's supposed to be doing. She's one of five (kids), and I think she, her mother started working this year, and she needs that extra attention. Whether it be positive, negative, whatever, it doesn't matter, just so she's gotten some kind of attention. It's funny, you finally get a handle on these kids and know where they're coming from and it makes it a little easier to work with them. (EE/STUD)

In several of the reflections coded (*EE*), a secondary focus emerged around *Instruction* (6/21). For example, Abby discussed how the pacing of instruction was influenced by the students.

> Normally I would teach it (the unit) . . . I had it set up instructionally the same way. What I think I had noted in here, depending on the class needs, it might be quicker (pace of instruction) . . . I really think depending on the class needs, and how the kids are doing, whether you speed it up or slow it down. (EE/INST)

Moreover, Abby recognized that she had made assumptions about students having certain experiences when in fact they had not had them. In the following episode coded EE/INST, Abby reflected how she might better help her students make partner selections for a paired reading activity.

> They started to get into their groups [pairs], it wasn't going too smoothly. I think they were uncertain as to just who they wanted to read it with. This is one thing I would change . . . the next time we did it, I had the pairs written up and they were on the board. So they knew exactly what they were doing. There I had made the assumption that they had already done things like that in the past and they hadn't. (EE/INST)

Explanations Based on Theory/Principle (ETP) and *Description* were equally represented in the transcript (18/57 or 32% of coded episodes). In most of the ETP episodes, *Instruction* emerged as the primary focus (12/18). For example, in the following episode coded ETP/INST, Abby reflected on the importance of providing many opportunities over the course of the school year for students to associate concepts learned by applying them in new ways.

> Right. And you are trying very hard to get it (concept) to relate somehow to make it start clicking. I do think it is a year-long process; it's something you have to keep going over and building on . . . You have to relate it to something you know in order for it to stick. . . . Throughout the year, you're constantly going back and pulling in and going back and pulling in more. That's like latitude and longitude (concepts introduced in this unit). I finally got to the section in the book where they introduce latitude and longitude. Well, we've been doing graphing all along with latitude and longitude. These kids are doing the same work as far as graphing goes as the sixth grade math class . . . And they love it. The concept has finally sunk in . . . Gridding the room, setting desks at certain points, having kids locate that, it helps to be able to visualize it. And then constantly pulling the vocabulary back in, so it's not like something we did two months ago and now we are moving on. I think it helps. (ETP/INST)

In other ETP/INST episodes, Abby observed that her lesson format, the way in which she introduced and closed the lesson, provided students with a comprehensive framework.

> Here, I used it (the closure activity) as a review, having them focus again. With something like this, you have to go back and have a focus of what you wanted them to get out of that. And keep going over and over and over it until it clicks. I think they did pretty well with it. Some of them had a handle on it and were starting to get an idea of labeling and looking at the map. (ETP/INST)

Abby reflected upon the ways in which she might change various aspects of the lesson in the following ETP/INST episode. She suggests improving the map labeling activity by coordinating the visual and auditory components of instruction.

> The reference map we used had different colors [than the directions for use of color the teacher was giving]. It had yellow in the book. I should have told them to use yellow. It's called being consistent. So there's some kind of continuity. So that when they're looking at one map, they may remember the colors and not so much the name. They eventually transfer them back and forth; maybe it might be a little easier transition. For some of the kids, it won't make any difference. They understand the concepts. (ETP/INST)

In the following ETP reflections, we gain a sense of Abby's commitment to know well her students, especially those with learning disabilities, and to address their needs. Shortly after the unit of instruction had been videotaped, three students with LD who had been receiving mainstream social studies instruction were pulled from the mainstream setting to be served full-time in a self-contained special education classroom. Abby cannot seem to find any justification for pulling these students from the mainstream setting, specifically for social studies instruction, and in this reflection alludes to the moral/ethical dimensions of teaching and learning.

> The problem with these kids [students with LD] for the first couple of years, they weren't forced into a situation where they had to apply themselves at all. And this was the first year, it was like trying to buck the system, they didn't want to do it, but I was bound and determined that they were going to do it, at least make an effort. At least make an effort and let me get a feel for what they were capable of doing and not capable. I've always kept the expectations of these kids high. And in the past, I've always found that the majority of them could perform to your expectations. And the more you kept raising them, the more they would strive to work toward them.
>
> They (students with LD) really have come a long way. I would have liked to have kept them all until the end of the year, just to see how much farther we could have taken them. They pulled them shortly after this . . . I was really upset that they did that. (ETP/INST)

As mentioned previously, nearly one third of Abby's reflective statements were coded as *Description* (32%). The predominant foci of these statements targeted *Students* (8) and *Content* (7). As the following episode illustrates, Abby was surprised at the difficulty her students had working in groups.

> It's surprising when I viewed this, too, how difficult it is. Because these kids, I found, could not work in groups. And even now, I'm just starting to get them into groups like this. We were broken up into groups and there would be five to six kids in each one of the groups. (D/STUD)

In summary, Abby's reflections illustrate how years of experience and knowledge of educational principles combined to anchor her teaching practices. When reflecting on instruction, Abby emphasizes the importance of providing numerous opportunities for students to actively apply their learning. Further, with students holding a major focus throughout the transcript, these reflections demonstrate Abby's efforts to know her students well. She acknowledges that her students directly influence the instruction. Abby's reflections provide evidence of the genuine sensitivity she exhibits toward meeting her students' needs.

Barbara

Barbara, a fourth grade teacher, was videotaped teaching a lesson which was about one third of the way into a Social Studies unit about mountains. Barbara's plans indicated that she hoped to engage the students in discussions and research which would focus on learning how mountains are a resource for the animals living there.

An analysis of the post-lesson interview indicated that the majority of Barbara's reflections (53%) were the type referred to as *Explanation Based on Experience or Personal Belief (EE)*. The predominant foci of these reflections were *Students* (9/18) and *Instruction* (6/18). For example,

> I wanted to do something with my wall map. . . They were pieces of paper, index cards that were cut. And I knew the names. I knew which animals were in all three mountain ranges. Then the kids just came up and put them on the map. So that's how that worked out. (EE/INST)

Barbara's experience has prepared her for some typical student actions to assistance from the teacher.

> . . . there were a couple of times I saw kids' hands up and I didn't get to them. And by the time I did get to them. . . the problem was solved and I know they're going to do that. You know, the majority of them are going to raise their hand as soon as they have the slightest question, without even looking at the book first. (EE/STUD)

In this episode of reflection, Barbara indicated that experience has shaped her current teaching practices.

> I don't talk as loud as I used to. My teacher voice changed . . . . I know that my student teacher yelled when she taught. And then the kids get louder. I'm more confident, I think, in myself than I was before . . . At one point in my teaching, I wasn't real comfortable with a lot of the things I taught . . . I think at one point I didn't organize as well before the lesson as I do now. (EE/INST)

When Barbara reflected on changes she would have made in the lesson, she also drew heavily on experiences and personal beliefs. The following comments relate to changes in grouping for the lesson, and how she should have managed feedback to students on work they had completed previously.

> I would group the kids instead of letting them group themselves. And that came from lack of knowing the class because it was early October and I didn't really know where they were. (EE/STUD)

> I was thinking about that yesterday. I don't know if I would pass the paper out first and then go over it with them . . . That might be better to do, have them look at it, the sheet they filled out . . . I probably would pass that out first, then review it. (EE/INST)

Barbara's reflections were not all based in personal experience, however. In nearly one-third of her reflections (32%), she drew upon her conceptualizations of professional theories. For example, Barbara reflected upon one component of lesson structure that had been stressed in the professional development workshop.

> I do a lot more of that [closure] than I ever did before. Now, in this lesson, I didn't because there really wasn't any way to close, and my closure in this lesson was letting the kids share their animals. (ETP/INST)

Barbara's comment on integration of content across subject areas is based on her application of professional theories.

> And a couple of things came up about the mountain ranges and the Appalachian Mountains and them traveling over them. It all fits. It all

> intertwines at some point and I think they saw it too. Because a lot of times they would come up to me and say "look what I found," and it'd be something we'd studied in Science we'd read about in Reading. And it sort of reinforced what we'd learned. (ETP/INST)

A small percentage of Barbara's reflective statements were coded *Description* (15%). The majority of these statements focused on *Students* (2) and *Instruction* (2). In the following episode, Barbara shared her positive impressions about the consistency among students within her classroom throughout the school year.

> I'm impressed with the way these kids were. It's them, it's the class too that makes your job easy or hard. They were easy, this group was so easy . . . And that's just the kind of class they were until the very last day. It amazes me. (D/STUD)

In summary, Barbara's reflections are primarily based on her experiences and beliefs, and they indicate a growing confidence in her instincts. Barbara states that her confidence has grown over the years from experience with students and familiarity with instructional materials and students. She notes that it had been helpful to her to have her practices validated through the workshops. Barbara felt that a good deal of the content of the videotaped lesson captured her usual approach and style.

> There were some things you might have seen that were so much a habit to me that they weren't unusual . . . They were just part of what I did every day.

Charles

Charles, a fifth grade teacher, was videotaped teaching a lesson near the conclusion of his unit on the Revolutionary War. The objectives of this particular lesson were to compare strengths of British and American sides and to define vocabulary related to the unit. The lesson began with a discussion of the vocabulary and continued with Charles placing a visual organizer on the board and asking students to think about the advantages each side had at the beginning of the war. This activity was followed by students reading aloud from the text and engaging in a discussion of the material.

In the post-lesson interview, Charles demonstrated all types of reflection (See Table 1), with the greatest percentage of episodes of the *Explanation Based on Theory/Principle* type (45%). These reflections focused primarily on *Instruction* (14/30), but Charles also discussed *Students* (7/26) and *Content* (8/26). Charles' comments related to instruction dealt with the importance of prior knowledge, the textbook, and strategies for teaching. For example, in several episodes, Charles discussed his beliefs about the importance of relating new information to prior knowledge, as well as his surprise at seeing how students attempted to make these connections.

> There were sections there where I was trying to bring their background and put them in the position. . . . I was surprised how much they did try to relate it, how they kept trying to pull back to now, and things they knew. . . . They were trying to take that information from history and make it part of their own life. (ETP/INST)

Charles also discussed his beliefs about strategies for teaching. For example, with respect to the use of oral reading, he says:

> [I use] volunteers. Even the kids who are in the LD class and have a hard time reading, they'll volunteer. I don't make anybody read if they don't want to; I don't put them on the spot, and I try to randomly just go around the room. Then I'll read some of it and if we're just sort of maybe spinning our wheels here, I'll say, "Here, you read the last couple of sections on your own and we'll get back together when you're finished." (ETP/INST)

Charles made definite statements about his use of discussion, explaining that he thought it was an important way to teach since it provided for active student involvement. In discussing the textbook, Charles talked about the importance of the textbook, but also the problems with using it.

> I think the textbook has . . . a very important place. I don't like a lot of the wording . . . I think it is a little too difficult for them to read on their own. I think a lot of times they, for lack of a better term, it's kind of high-falutin. They don't seem to just write to a fifth grader. (ETP/INST)

A final example of a reflection coded ETP includes an episode in which Charles demonstrated his willingness to address ethical or value-laden issues, and his attempts to present both sides. The episode dealt with the involvement of the United States in the Vietnam War, essentially asking the question, "Why were we there?" Charles labeled this interaction as a "digression" but justified it by saying that students do bring up issues that relate to current events or their own knowledge.

> I don't sugar coat any of this either [controversial issues]. . . . I don't make it sound like the U.S. was always right and correct in what they did. It's impossible to keep our values out of it completely . . . I try to step back from it and say, 'There were two sides to it,' I do try to give both sides or all sides as much as possible. (ETP/CON)

A second type of reflective statement was coded almost as frequently as ETP. These were reflective statements that were coded as *Explanations Based on Experience and Personal Belief (EE)* (33%). Again, during the EE episodes, Charles discussed *Instruction, Content*, and *Students*. Throughout the interview, Charles kept coming back to the amount of information he had in the lesson, indicating "I was surprised at how much material I actually had. . . . and I probably should have broken it apart." He also came back to the importance of helping students make connections. "I try to think of when I was [age] 10. Whether this would have made any sense or not."

Reliance on experience was also coded when Charles talked about the involvement of his students. He was sensitive to the fact that students were interested in the lesson, raised their hands, and asked questions of him.

Charles also discussed one change that he would make, again relying on his beliefs and experience in this description of his use of a graphic organizer. Charles noted that the graphic organizer he placed on the board was developed in such a way as to generate only the advantages to the British in the Revolutionary War.

> The one thing I thought was not real good was what I was doing at the board. I did advantages and disadvantages, but what I really never got were disadvantages, because it was all basically the British. . . . It probably would have been better to break it down into American and British and . . . (EE/INST)

Some of Charles' reflective statements were coded as *Description* (22%). He pointed out students who were labeled as learning disabled and mentioned how involved they were in the lesson. "I thought they were very interested and in this discussion here, there were a wide range of kids who were involved." (D/STUD)

In summary, Charles' reflections derive from both his strong knowledge base and his teaching experience. In thinking about how and what he teaches, Charles relies on strong beliefs about what he does and can justify his teaching style and format using both theory and experience. He is aware not only of content issues, but of instructional and student issues, as well. At the same time, because management is not of concern, Charles spends almost no time discussing it.

David

David, a sixth grade teacher, taught a unit of social studies instruction on Ancient Egypt. The videotaped lesson occurred approximately midway through the unit, the fourth day of a 7-day unit. The main objective for this day's lesson was to have students demonstrate, through a writing activity, their understanding of the Egyptians' belief in the afterlife. David, attired in Mummy-like gauze, described the Egyptian concept of the afterlife as a rationale for the practice of mummification, an essential part of the Egyptian funeral ceremony. Then students used the textbook to engage in a paired reading activity on the topic of "The Book of the Dead." Following this activity, students were asked to act as a "scribe" and write entries into a "Book of the Dead" for a mythical pharaoh. Students were to include such things as advice on talking to the gods, and a list of the wrongs that this pharaoh might have committed or avoided.

During the interview, the majority of David's reflections were coded as *Explanation Based on Theory/Principle (ETP)* (42%) (See Table 1). The majority of these reflections (13/23) as well as his reflections in general (49%) focused on *Instruction*. David showed a keen awareness of the importance of tapping into prior knowledge of students and using graphic organizers and the text as instructional tools.

> . . . Something we had done on Day 3 [of the unit of instruction], we had brainstormed. When we introduced the unit, I asked them to brainstorm terms. Anything that they thought that was associated with Egypt. We brainstormed, had a "scribe" write them all down on the board, then from that list, they categorized them on that topical net [graphic organizer]. (ETP/INST)

> I wanted the students to understand the process of mummification and how important it was to the Egyptians in terms of their beliefs. How the process is very involved and very ritualistic and complicated. Obviously the tie in was to their [the students'] religious beliefs. (ETP/INST)

For David, tapping into students' prior knowledge and experiences was an essential way to build on new concepts introduced.

> OK, they were associating a familiar activity. A lot of these kids from the country hunt. They were associating that, the act of gutting to what happens with the mummy . . . I picked that out because they were taking association or prior knowledge. They have the prior knowledge of a particular act that they could associate to this process in ancient Egypt. And there were about four different places (during the lesson) where they said hey that's like . . . and they associated the act of what is being described here. (ETP/INST)

Some of the reflections coded ETP showed David's emphasis on students having frequent learning experiences in which they feel they are contributors in the learning process. Furthermore, David recognized ways he could enhance students' comprehension when teaching these concepts in the future.

> All these references to beef jerky or deer jerky, I should have gotten some jerky to pass around. The feeling of it, they could tactically feel the dry flesh and just what it feels like . . . and that is what I should have done. (ETP/INST)

In other reflections coded ETP, David demonstrated how the process of engaging in reflection led to questions regarding the extent to which students grasped newly introduced concepts.

> I wrote down here that my explanations were poor. Not enough examples, it was too foreign . . . It was a task that was unfamiliar to them, listing things the pharaoh should have done or did right. (ETP/INST)
>
> Again, something we have already brought up. The notion that strong point may nest in a student's brain and that does not allow them to think about the subsequent points. When I saw that I thought that is something we must address because that is extremely important. That raised a big question in my mind about my own practices. I tend to make strong demonstrations. They work for that one point that I am making but does it in some way detract from some points to come? I don't know. (ETP/INST)

In one ETP reflective statement, David was forthright in acknowledging an inappropriate response he made to a student in his class. He openly shared his lack of sensitivity toward the student.

> I made a cheap shot. It was one of these things that I said something and thought gee I shouldn't have said that . . . Setting a bad example, I'm not sure that happens more frequently than teachers would like to believe. It is like a rim shot, it was harmless and it wasn't anything that disrupted the class but my response maybe should have been maybe to chuckle or pass it by, but instead I made him (the student) the butt of a joke. Now this is a kid who is a very bright student, now if a lower-ended kid would have said something like that, I hope that I would have been more sensitive. (ETP/STUD)

David felt that his positioning within the classroom had been a successful tool as illustrated in this episode which focused on instruction and classroom management.

> What I wanted to point out and then the camera pulled away, this configuration enables me to move right up next to the majority of the students in class. This is intentional. This is purposeful. I intentionally went into the audience so to speak when I was changing the focus. We were going from the mummification to the writing assignment to the Book of the Dead. The movement within the room again was an attempt to say was here I am . . . something important is happening. (ETP/INST) (ETP/MANAGE)

Explanations Based on Experience and Personal Belief (EE) were the second most frequent type of reflection in David's transcript (33%). Again, most of these reflective episodes focused on *Instruction* (10/18). In these episodes, David reflects a belief that experience taught him that timing is a key factor in the success of presenting new information.

> With that social studies class, the bottom-line fight is that it is at the end of the day. It is the last class of the day and they are pretty much rung out by that time. So a passive writing exercise of a workbook page is something that doesn't, I don't think it makes an impact . . . So I think that with the art of teaching, I think timing is an important factor. (EE/INST)

Also, David felt that he would change the order and type of activities the next time he taught this particular lesson, based on this experience.

> It wasn't a simply respond to the question type thing. It was a create assignment, and again like I said, looking back on it, it was a poor choice of

> assignment and a poor choice of topics for that particular time of day. I don't think it worked. I don't think it worked well at all. (EE/INST)

About one fourth (25%) of David's reflective statements were descriptive in nature (D). *Content* (6) and *Instruction* (4) surfaced as the major foci of these reflections. In the following episode, David describes in a humorous manner how he dressed as the man who discovered the tomb of King Tut.

> I was Howard Carter. I was the guy who discovered King Tut's tomb. I wore a pair of khaki pants and khaki shirt and a pith helmet with a little black tie. I looked like the guy who opened the coffin in the Boris Karloff movie. (D/INST)

In summary, David's reflections derived from his knowledge of educational principles as well as from his extensive experience. The interview transcript provides insights into his sensitivity toward students and his keen interest in keeping students actively involved in the learning process. Finally, from these reflections, we gain a sense of David's confidence as a teacher; he is comfortable in raising questions regarding the relevance and appropriateness of his own teaching practices.

Themes

In this section, we discuss themes across all four teachers. These include a summary of types of reflection and foci, a summary of teachers' ability to identify strengths of their lessons or changes they would make, and an analysis of the relationship between their statements and the content of the initial workshop. We also describe the teachers' responses to this opportunity to reflect.

Reflections/Focus. All four teachers most frequently used a combination of *Explanations Based on Experience and Personal Belief* and *Explanations Based on Theory and Principle* as they reflected on the videotaped lesson each had taught. There was evidence of strong beliefs regarding instruction, as teachers described their styles of teaching and the strategies they used as they taught. These four teachers were also sensitive to their students, able to identify students who were involved as well as those who were having difficulty attending. They were sensitive to the context in which they taught, discussing factors such as prior experience of students, both personal and educational, time of day and year. Content was discussed more frequently by the two teachers who taught at higher grade levels and both these teachers seemed as concerned about the outcomes of instruction as they did about the process. In contrast, the two teachers who taught earlier grades appeared to be more concerned with the processes of instruction than with content. Three teachers also described situations in which ethical or moral considerations were primary.

Identification of Strengths. All four teachers could identify strengths of their lessons; most of these strengths centered around instructional practices. Abby and Barbara were pleased with the lesson framework they had used. (Teachers in the workshops had been taught to open each lesson with a rationale for the lesson so that students would know what, how, and why they were doing a task, and to close each lesson in the same way.) David was pleased with the connections that he had made between lesson content and prior knowledge. He was also pleased with the fact that students felt as though they were contributors in the learning process.

> I noticed my own comment to the student [student offered a response and teacher commented 'Gee, I hadn't thought about it like that']. What I wrote here on these notations is that the student feels like he is a contributor . . . That kind of response can make the student feel like he is not just a receiver but he is a contributor also. (David)

Charles was pleased with the discussion format and the active involvement of the students.

> I thought the discussion was very well laid out and beneficial. . . . I like that kind of a format. . . . I think you get a very good feel for what they understand when you're doing it with them like that. (Charles)

He was also able to explain in an articulate fashion the strategies that he used to enhance understanding, including "verbal" discussion, oral reading, and relating content to prior knowledge (what he called "personalizing instruction").

Identification of proposed changes. All four teachers identified changes that they would make were they to teach the lesson again. The focus was mostly instructional, with several references to students. Abby indicated that she would provide more guidance to students in making partner selections, solicit more information to increase students' level of active involvement, and improve the map labeling activity by coordinating the visual and auditory components of instruction. Barbara thought she would change the grouping in the lesson:

> I would group the kids instead of letting them group themselves. And that came from lack of knowing the class because it was early October and I didn't really know where they were. (Barbara)

As indicated previously, Charles was disappointed in how he used the visual organizer and would change that structure. He was also concerned about the amount of information he had tried to cover in the lesson.

> The lesson . . . it was a little long. . . a lot of information. Probably, I may have even broken it apart into two lessons . . . used the summary of all the information we had before. (Charles)

David related a number of ways in which he would change the lesson: attending to different modes of student learning, providing more opportunities for active involvement, and changing the order and type of activities presented in this lesson:

> My perception of the lesson was more or less that it was a passive one. The student was a receiver and I was a lecturer. You never know whether there is going to be a lot of student response or not. I don't like a whole lesson full of passivity (among students). (David)

Relationship to the Workshop. For all four teachers, we could find evidence in their reflections of topics that had been discussed in the previous workshops. Charles and David were attentive to the limitations and the potential of the textbook. Abby and Barbara used the lesson framework that had been described and were pleased with its effect. And all four teachers put a great deal of emphasis on the importance of helping students relate their new learning to prior knowledge and on the importance of active involvement of students.

The Power of Reflection. One of the significant findings of this study was that these experienced teachers not only could reflect on many different dimensions of their lesson, but also appreciated the opportunity to do it. Moreover, it was obvious from their comments that these teachers did not usually take the time to think about or reflect on their practices. The task of viewing and reviewing a lesson, even long after that lesson had been taught, provided these teachers with new insights and surprises. The teachers recognized erroneous assumptions about the prior learning of students and also some aspects of their own teaching that perhaps could be improved. The opportunity for reflective thinking about teaching, for experienced teachers, appears to be something that can assist them in becoming more thoughtful about why, what, and how they plan and conduct instruction. Perhaps David said it best "This process of us looking at this is good; you know it is very good for me too, really."

Conclusions

These experienced elementary teachers had strong beliefs about their instructional practices and about the students they taught. Some beliefs were based on the many experiences they have had throughout the years; other beliefs were based on solid theoretical principles. The statements these teachers made in discussing theoretical principles seemed to indicate that these teachers had a good sense of their role as professionals and of the pedagogical theories supporting their work. It appeared that their many years of experience had helped them integrate various activities based upon pedagogical theories into their classroom practices. They could speak comfortably and with confidence about what they were doing and why. In fact, the statements of each of these teachers reflected a strong sense of voice and a trust in their own experiences and knowledge as a basis for decision-making. As Canning (1991) states, "the taking on of an 'I' voice was one of the achievements of the reflection process." (p. 19). The teachers in this study were eager to present their points of view and raise questions about their own performances in the classroom. These teachers, who were volunteers, may have been more intellectually curious about their teaching performance, and perhaps more professionally oriented than experienced teachers as a whole. Certainly, their ability to reflect on their performance was a good indication of their professional interest in and concern about themselves as teachers.

The foci on *instruction* and *students* rather than *content* may be indicative of the grade levels at which these teachers taught. Overall these teachers appeared to be more attuned to how they taught and to the ways in which students could be actively involved than to the content of their teaching. It may be that they are more accepting of the curriculum (content) as it is; or, it may be that the content is not as important to them as helping students learn to learn.

The lack of emphasis on management appears to reflect the fact that these teachers had organized their classrooms and their instruction in ways that reduced or minimized any difficulties with classroom management. Certainly, the lack of concern about management provided more opportunity for teachers to reflect about other aspects of their teaching.

All four teachers incorporated ideas from the workshops into their teaching and they could identify these elements as they reviewed their videotape. As we read about school reform and restructuring, there is a great deal of discussion about how to provide staff development for experienced teachers and how to involve teachers in the design of such programs. In this study, teachers were provided with an intensive one-week experience in the summer, assisted individually with the design of a unit that incorporated principles promoted in the workshops, and then observed as they taught the unit. The interviews with the teachers, even long after the lessons were taught, indicated that they were aware of the principles that had been promoted as part of the workshops, and as importantly, could identify instances where they had incorporated these principles into their teaching. Whether these teachers continued to use these principles is a question that we cannot answer; nevertheless, the fact that they could relate in an articulate manner their responses and those of the students to the principles they had learned appears to be a positive indication that the workshops did have a positive influence.

It would appear that these four experienced elementary teachers gained much from the opportunity to participate in a review and analysis of a taped lesson. All four teachers could identify what they saw as strengths of their teaching and possible changes they would make in the future. They also expressed surprise at what they learned by watching the lesson and reflecting about their teaching. Each teacher was able to articulate in a professional manner many different dimensions of his/her teaching. Such responses lead us to the conclusion that the opportunity to reflect may be an important experience not only for beginning teachers but for those who have been teaching for many years. A structured set of opportunities for "reflection" may be as important to staff development as any information provided to teachers.

Further, the reflections of teachers can help researchers gain a better understanding of what happens in the classroom. As Meek (1991) states, "teachers have so much knowledge about how classrooms work and about kids' lives in classrooms, and that knowledge on the whole is untapped and known only to the person who holds it. . . . it's important to help teachers appreciate that they have knowledge worth taking seriously" (pp. 33-34).

Future research might focus on practical ways to use reflection to achieve instructional change within existing school contexts. Further, opportunity for teachers to see and then think about how they have included new ideas in their instructional repertoire may be an essential part of any restructuring or reform movement.

References

Adler, S. (1994). Reflective practice and teacher education. In E. W. Ross (Ed.), *Reflection practice in social studies, Bulletin 88* (pp. 51-58). Washington: National Council for the Social Studies. (ED 373 014)

Askling, B., & Almen, E. (1995, April). *The reflective turn in teacher education: Possibilities and limitations of an implementation in teacher education and for forming "reflective practitioners."* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Canning, C. (1991). What teachers say about reflection. *Educational Leadership, 48*(6), 18-21.

Denton, J. J., & Peters, W. H. (1988). Development of reflective thinking skills about pedagogy. In H. C. Waxman (Ed.), *Images of reflection in teacher education* (pp. 40-41). Summaries of papers presented at a National Conference on Reflective Inquiry in Teacher Education. Houston, TX. (ERIC Document Reproduction Service No. ED 316 542)

Evans, C. (1991). Support for teachers studying their own work. *Educational Leadership, 48*(6), 11-13.

Griffin, G. A. (1991). Interactive staff development: Using what we know. In A. Lieberman & L. Miller (Eds.), *Staff Development for Education in the '90's.* (2nd ed) (pp. 243-258). New York and London: Teachers College Press, Columbia University.

Hayes, L. F., & Ross, D. D. (1988, October). *Trust versus control: The impact of school leadership on teacher reflection.* Paper presented at the Florida Conference on Reflective Inquiry: Contexts and Assessment, Orlando, FL. (ERIC Document Reproduction Service No. ED 307 265)

Holly, M. L. (1983). *Teacher reflection on classroom life: an empirical base for professional development.* Progress report #3. Kent, OH: Kent State University. (ERIC Document Reproduction Service No. ED 243 865)

Houston, W. (1988). *Reflecting on reflection in teacher education.* Paper presented at the National Conference on Reflective Inquiry in Teacher Education. Houston. (ERIC Document ED 316 542)

Killion, J. P., & Todnem, G. R. (1991). A process for personal theory building. *Educational Leadership, 48*(6), 14-16.

Knapp, C. E. (1992). *Lasting lessons: A teacher's guide to reflecting on experience.* Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools Appalachian Education Laboratory.

Marshall, H. H. (1990). Metaphor as an instructional tool in encouraging student teacher reflection. *Theory into Practice, 29*(2), 128-132.

Meek, A. (1991). On thinking about teaching: A conversation with Eleanor Duckworth. *Educational Leadership, 48*(6), 30-34.

Moore, J. R., Mintz, S. L., & Biermann, M. J. (1988). Reflectivity: the Edsel of education? In H. C. Waxman (Ed.), *Images of reflection in teacher education* (pp. 45-46). Summaries of papers presented at a National Conference on Reflective Inquiry in Teacher Education. Houston, TX. (ERIC Document Reproduction Service No. ED 316 542)

Munby, H., & Russell, T. (1994). The authority of experience in learning to teach: Messages from a Physics methods class. *Journal of Teacher Education, 45*(2), 86-95.

Oja, S. N. (1991). Adult development: Insights on staff development. In A. Lieberman, and L. Miller, (Eds.), *Staff development for education in the '90's* (2nd ed.). (pp. 37-60). New York and London: Teacher's College, Columbia University.

Pugach, M. C., & Johnson, L. F. (1988). Promoting teacher reflection through structured dialogue. In H. C. Waxman, (Ed.), *Images of reflection in teacher education* (pp. 30-31). Summaries of papers presented at a National Conference on Reflective Inquiry in Teacher Education. Houston, TX.

Richert, A. E. (1991). Using teacher cases for reflection and enhanced understanding. In A. Lieberman, & L. Miller, (Eds.), *Staff development for education in the '90's* (2nd ed.) (pp. 113-132). New York and London: Teacher's College, Columbia University.

Ross, D. D. (1988). Reflective teaching: Meaning and implications for preservice teacher educators. In E. C. Waxman, (Ed.), *Images of reflection in teacher education* (pp. 25-26). Summaries of papers presented at a National Conference on Reflective Inquiry in Teacher Education. Houston, TX.

Ross, E. W. (1994). (Ed.) *Reflective practice in social studies*. Washington: National Council for the Social Studies. (ED 373 014)

Russell, T. (1993). Critical attributes of a reflective teacher: Is agreement possible? In J. Calderhead, & P. Gates, (Eds.), *Conceptualizing reflection in teacher development* (pp. 144-153). London: The Falmer Press.

Schon, D. (1983). *The reflective practitioner*. New York: Basic Books.

Seidel, J. (1988). *The Ethnograph*. Littleton, CO: Qualis Research Associates.

Smith, R.W. (1991, April). *Obstacles to student teacher reflection: The role of prior school experience as a barrier to teacher development*. Paper presented at Annual Meeting of the American Education Research Association. Chicago, IL. (ERIC Document Reproduction Service No. ED 336 352)

Sparks-Langer, G. M., & Colton, A. (1991). Synthesis of research on teachers' reflective thinking. *Educational Leadership, 48*(6), 37-44.

Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., & Starko, A. (1992). Reflective pedagogical thinking: How can we promote it and measure it? *Journal of Teacher Education, 41*(4), 23-32.

Thornton, S .J. (1994): Perspectives on reflective practice in social studies education. In E. W. Ross (Ed.), *Reflective practice in social studies* (pp. 5-11). Washington: National Council for the Social Studies, Bulletin Number 88.

Van Manen, M. (1977). Linking ways of knowing with ways of being practical. *Curriculum Inquiry, 6*(3), 205-228.

Wildman, T. M., & Niles, J. A. (1987). Reflective teachers: Tensions between abstractions and realities. *Journal of Teacher Education, 38*(4), 25-31.

Zeichner, K. M. (1990, November). *Educational and social commitments in reflective teacher education programs*. Proceedings of the National Forum of the Association of Independent Liberal Arts Colleges for Teacher Education, Milwaukee, WI. (ERIC Document Service No. ED 344 855)

Zeichner, K. M., & Liston, D. P. (1987). Teaching student teachers to reflect. *Harvard Educational Review, 57*(10), 23-48.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 39-47

Teacher Perception of Kentucky Elementary Principal Leadership Effectiveness and School-Based Council Meeting Effectiveness

Patricia Lindauer
Hardin County Schools, Elizabethtown, Kentucky

Garth Petrie and Michael Richardson
Georgia Southern University

This study addressed the current trend of school governance by school-based councils. Its major purpose was to determine if there was a relationship between teachers' ratings of a) principal's leadership effectiveness, and b) school-based council effectiveness. Returns provided data concerning 320 elementary teachers and 19 elementary principals. The Pearson R revealed a moderate positive relationship (r = .62) between teachers' ratings of principal's leadership effectiveness and council meetings' effectiveness. However, the Pearson R showed no significant relationship between teachers' ratings of principal leadership range and council meetings' effectiveness. The t test indicated a significant difference (t = 2.19) between council teachers' ratings and non-council teachers' ratings of principal's leadership effectiveness.

Introduction

Leadership is one of the most researched areas in the behavioral sciences, yet basic problems remain–one of which is differing definitions of leadership (Lunenberg & Ornstein, 1996). The literature is filled with leadership theories and studies which attempt to define leadership. Koontz and O'Donnel (1959) stated that "leadership is influencing people to follow in the achievement of a common goal" (p. 453). Terry (1960) defined leadership as the "activity of influencing people to strive willingly for group objectives" (p. 493). Hersey and Blanchard (1988) defined leadership as the "process of influencing the activities of an individual or group in efforts toward goal achievement in a given situation" (p.85). Regardless of the definition chosen, leadership definitions seem to follow a common theme–influencing people to achieve common goals in a given situation.

In 1983 the National Commission on Excellence in Education released its report, *A Nation at Risk*, which reported in detail the shortcomings and decline of the American educational system (Bell, 1983; Hammond, 1990; Sevener, 1991). It was followed by recommendations by the Holmes Group (1986) and by the Carnegie Task Force on Teaching as a Profession (Carnegie Forum on Education and the Economy, 1986) that decisions should be made at the level at which they are implemented. In 1989 President George Bush met with governors of the 50 states, and they developed *America 2000*, a strategy to move the nation toward achieving national educational excellence. One of the 15 accountability strategies addressed in *America 2000* was establishing Governors' Academies so that principals and other leaders would be able to make their schools better and more accountable (Lieberman, 1988a; U.S. Department of Education, 1991).

Principal Leadership

During the last five decades there have been numerous studies concerning leadership effectiveness and style. Research suggested that the effectiveness of a given leader could be improved through a better understanding of leadership. Effectiveness, however, like leadership, represented both principals' and teachers' biases, ideas, and values, and was an "artificial construct" in each of their minds (Cross, 1981). Individuals' basic motivation and need structure would be reflected by their leadership styles (Fiedler, 1967). To some extent, then "leadership is making happen what you believe in" (Barth, 1990, p. 515).

Patricia Lindauer is the Director of Program Planning and Evaluation for the Hardin County Schools in Hardin County, Kentucky. She currently is on leave from her district and serving as an assistant professor at Georgia Southern University. Garth Petrie is a professor of educational leadership at Georgia Southern University in Statesboro, Georgia. Michael D. Richardson is a professor of educational leadership at Georgia Southern University in Statesboro, Georgia, and is coordinator of the doctoral program. Correspondence regarding this manuscript should be addressed to Patricia Lindauer, College of Education, Georgia Southern University, P. O. Box 8131, Statesboro, GA 30460-8131 or by e-mail at plindaue@gsvms2.cc.gasou.edu.

Each school varied in its own combination of character, style, and substance. How effective the principal was in each new role was largely determined by his/her leadership methods. However, the leadership style that was effective in one setting was not necessarily effective in another (Kloph, Scheldon, & Brennan, 1982; Manasse, 1982; Sexton & Switzer, 1977; Thomas, 1977). Hersey and Blanchard (1988) believed that the major attribute which set the successful organization apart from the unsuccessful organization was dynamic and effective leadership. The leadership style currently in vogue is the transactional or situational leadership style. Roles of leaders and needs of followers must be congruent in this style; the leader must blend the role and subordinate needs with the situation (Hoy & Miskel, 1993; Pascarella & Lunenburg, 1988). The situational leadership theory that formed the basis for Hersey and Blanchard's Tri-Dimensional Leader Effectiveness Model (1987) was based on the amount of direction and socioemotional support a leader must provide given the situation and the level of maturity of the followers in relation to a specific task. According to this theory, there was no one best way to influence people.

A review of the literature led the researchers to the conclusion that effectiveness in leadership is a function of the leader, follower, and situation. The leader should adapt to the situation and the needs of the followers to be effective. The goal of the leader is to provide the necessary leadership behavior while simultaneously helping the group mature and assume more of the leadership itself. Vroom (1976) lends support to the situational theory in his contingency approach to leadership. Hoy and Miskel (1993) reported that the assumption behind the Vroom and Yetton contingency model was that situational variables which interacted with personal attributes of the leader resulted in leader behavior that impacted the effectiveness of the organization.

The literature also suggested three other important factors in leadership. The first was perception, the second was relevance, and the third was communication. Calder (1977) and Pfeffer (1977) began the discussion when they made the assumption that leadership existed in people's minds rather than representing an objective reality. Mitchell and Tucker (1992) supported this contention by claiming that leadership was a way of thinking and feeling about ourselves, our jobs, and the nature of the educational process more than it was a matter of aggressive action. All four authors suggested that leadership was a perception, a way of thinking and feeling about oneself and others.

Hammersely (1990) developed the concept of relevance by maintaining that how we describe an object depends not just on decisions about what we believe to be true but also on the judgments we make about the relevance of those actions viewed. The individual's or group's view of leadership, then, was dependent on that individual's or group's values, feelings, and assumptions at a given time and place. In other words, perception was the use of the senses to collect data and the background of our experiences to interpret those data. These perceptions developed over time and in given circumstances are fixed and influence future perceptions and actions.

According to Blase (1987) and Whaley (1994), teachers' perceptions of administrative consideration affected their job, with consideration being the extent to which they perceived the administrator engaging in two-way communications, listening and giving feedback. McClelland (1965) supported this idea with his achievement-need motivation theory. He suggested that all people have three basic needs–affiliation, achievement, and power. He believed all three of these needs came into play in some way in motivating an individual's behavior.

When principals worked with school-based councils, this leadership perspective took place in a live setting. The research literature on school councils, though limited, seemed to indicate that teachers' perceptions of the effectiveness of the school-based council meetings would positively influence the success of the council in implementing change (Bahrenfuss, 1992; Bergman, 1992; Mullen, Symons, Hu, & Salas, 1989). Brandt (1982) concluded that effective schools have effective principals, bringing the researchers to conclude that successful site-based schools need principals who can effectively develop and conduct council meetings in such a way as to make them more effective also. In other words, teachers' perceptions of effective principals in school-based leadership situations were connected to success in school-based decision-making situations.

School-Based Decision-Making (SBDM)

Education has been under one of its most severe attacks since the 1983 publication of *A Nation at Risk*. This reform package placed the blame for a mediocre educational system on the teachers and was intent on making schooling more demanding (Hansen, 1991; Lieberman, 1988a; Timar & Kirp, 1989). The second generation of reports were the Holmes Group Report (1986) and the Carnegie Task Force on Teaching as a Profession (Carnegie Forum on Education and the Economy, 1986), which said the problem was with the structure of the organization and not with the teachers. These two groups called on school districts to give teachers a greater voice in school decisions. Following this lead, the National Governors' Association in 1987 and the National Education Association in 1988 issued reports emphasizing allowing the participation of teachers in decision-making at the school site (Conley & Bacharach, 1990; Lieberman, 1988a; Meadows, 1990).

The more current buzz-phrase and trend in education circles today seems to be school-based decision-making (Mentell, 1993). However, as long ago as 1946, Alice Meil, an educational researcher, called for school-based management councils (Raywid, 1990). Early multi-classroom schools were managed by a head teacher and classroom teachers. However, after World War II the growth of teachers' unions and the proliferation of school administrators put the two groups in conflict. This caused the schools to become more centralized and bureaucratic (Fitch, 1991; Raywid, 1990; Taylor & Levine, 1991). School-based decision making, though not a new idea, is a complex process. The 1990s perception of school restructuring zeroed-in on "individual schools as an essential element to change and reform" and perceptions of school leadership as a shared endeavor (Rothberg & Hill, 1992).

Herbert Klausmeier and a team of educators at the University of Wisconsin, Madison, developed a school improvement plan known as Individually Guided Education. This group of education pioneers realized the importance of teachers being involved in the day-to-day decisions, and these decisions needed to take place at the schools. However, these improvements required the support of the principal and the central office administration (Fitch, 1991; Taylor & Levine, 1991).

Along the same line, John Goodlad, 18 years later, argued school-based management would make schools more flexible, accountable, productive, cost effective, and efficient (Fitch, 1991; Taylor & Levine, 1991). David (1989) stated this succinctly when she said that under the school-based management structure professional responsibility replaced bureaucratic regulation.

As a reform strategy, school-based decision-making required two changes. First, schools had to be organized and institutionalized differently. Second, schools had to be held accountable. Since school-based decision-making has not been in operation very long, there is little evidence of its impact (Taylor & Levine, 1991). Nevertheless, many current educational reforms include recommendations that local districts adopt some form of school-based management.

Edmonton, Alberta, Canada schools, according to experts, were one of the best examples and have gone further than many of the school-based decision-making schools. The Edmonton Board of Education set the goals, and the schools had the option to meet these goals in ways most conducive to their students (Dreyfuss, 1988; O'Neil, 1996).

In the early 1970s, the Dade County, Florida school boards approved certain regulations for shifting responsibility from the central office to the individual schools (Gomez, 1989; Lieberman, 1988b). This reform action was initiated by the superintendent and the teacher union, but the school board was not yet ready to implement school-based decision-making fully (Gomez, 1989; Timar, 1990). However, because of an attitude shift nationwide, in 1986-1987 all Dade County schools were asked to submit a proposal to participate in a school-based decision-making pilot program.

In the Hammond, Indiana system, the ultimate power to change was in the hands of the educators who worked in the schools. Schools were the front lines of change (Sirotnik & Clark, 1988). This district demonstrated a deep understanding of these principles when in 1985 it began the School Improvement Process (SIP). Teachers, administrators, parents, and students comprised committees which made decisions concerning curriculum, instruction, staffing, professional development, discipline, scheduling, etc. (Casner-Lotto, 1988).

The Los Angeles school-based decision-making model was a product of the teachers' union and the school board, which established Local School Leadership Councils (LSLC) in 1989. These councils, composed of 16 members with the majority being teachers, made decisions concerning areas such as scheduling and allocation of discretionary funds (Afsahi, 1990).

Chicago-style school reform was a product of the Illinois School Reform Act of 1985 and was legislated in the Chicago School Reform Act of 1988 (Fitch, 1991; Hansen, 1991; Walberg & Niemiec, 1994). These reforms, implemented in 1989, made a major difference in the way Chicago school-based management was handled in that the eleven member school-based council was controlled by parents/community members. These councils in Chicago, called Local School Improvement Councils, had genuine decision-making power. The one major difference in Chicago's approach to council decision-making was in the council's power to hire and fire principals (Bryk, Easton, Kerbow, Rollow, & Sebring, 1994; Fitch, 1991; Hansen, 1991; Rist, 1990).

In 1990 the Boston Teachers' Union negotiated a teacher contract to begin a second try toward school-based management. As a result, 34 schools had formed school site councils (SSCs) by 1993. Their focus was a collaborative approach with a group of five local educational and business organizations. The goal of this collaborative group was to develop school site council capacity so the councils could effectively manage the educational direction of their school, and to assist the central office personnel in developing supportive procedures and facilitator skills (Gleason, Donohue, & Leader, 1996).

Canton Middle School in Maryland turned to school-based management literally to save itself. In 1991 this urban school sought to empower teachers through a collaborative school-based management approach, including a teacher-designed curriculum implemented to help students develop their intellectual, social, emotional, and physical capacities (Spilman, 1996).

Statement of the Problem

There have been many studies and surveys examining principal leadership style. However, there has only been a limited amount of research correlating the relationship between the principal's leadership effectiveness and effectiveness of school-based decision-making council meetings. The present study was designed to examine the relationship between the principal's leadership effectiveness and the effectiveness of the school-based council meetings as perceived by the teachers in that school.

Research Questions

The major research question that guided this study was:

1. Is there a relationship between the teachers' ratings of principal's leadership effectiveness and the teachers' ratings of school-based decision making council meetings' effectiveness?

Other research questions which served as guides to the study were:

2. Is there a difference between teacher council members' and teacher non-council members' ratings of the principal's leadership effectiveness?
3. Is there a relationship between the teachers' rating of principal's leadership style range and the teachers' rating of the effectiveness of the council meetings?

Method

Subjects

For this study, data were collected from teachers and principals in a stratified randomly selected cluster sample of elementary schools in Kentucky. Twenty-six schools from 6 of the 7 congressional districts agreed to participate in the study, with 23 schools returning the information. Of these 23, 19 or 82% returned usable data, including 320 teachers and 19 principals. All four of the unusable returns had incomplete data.

The teacher sample included 85.9% females and 13.0% male respondents. The age range was from 22 to 60 plus years, with the mean teacher age of the respondents being 50 and the standard deviation being 8.94. The largest group, 40.6%, fell in the 41 to 50 age range. White respondents represented the largest category with 95.3%. African Americans represented the next largest group with 4.1%. According to the U. S. Department of Commerce's 1990 United States Census report, 7% of the total population in Kentucky was African American, with between 2 and 4% of the teacher population listed as African American, indicating the sample fairly represented the Kentucky teacher population. Only 54.4% of the teachers surveyed had obtained a Master's Degree, while 16.2% held the Bachelor's Degree.

Demographic data collected from the 19 principals showed 68.4% male as compared to 31.6% female. The majority of those surveyed, 52.6%, were in the age range of 41-50, with the mean principal age being 43.7 with a standard deviation of 7.31. The smallest percent of principal respondents, 15.8, fell in the 51 to 60 age range. Eighteen of the 19 principals were white, giving a percentage of 94.7. All the respondents had obtained a Master's Degree, with 68.4% holding a Rank I certification, which is 30 hours above the Master's Degree.

Instruments

The Leader Effectiveness and Adaptability Description (LEAD/Self and Other), developed by Hersey and Blanchard (1988), and Meetings, developed by Miles (1968, in Lake), were the instruments used in this study to measure teachers' ratings of principal's effectiveness and school-based council meetings' effectiveness. The LEAD/Self and Other were used to assess the perceived leadership behavior of the principals with whom the teachers worked and the principal's own perception of his/her leadership behavior. The LEAD/Self and Other questionnaires were developed to measure three aspects of leader behavior: style, style range (flexibility), and style adaptability (effectiveness). Style and style range were determined by four different style scores (quadrants), while style adaptability was determined by one normative score. Style range referred to the extent to which the principal's style could be varied while style adaptability indicated the degree to which changes in styles were appropriate to the readiness level of the teachers involved in the different situations. The LEAD/Self and Other instruments contained twelve leadership situations from which respondents were asked to decide between four alternative solutions. The tasks are described in terms of the degree of task and/or relationship behavior: high task/high relationship, high task/low relationship, low task/high relationship, and low task/low relationship behavior. Each of the situations also described the different maturity levels and styles of the followers: high maturity, high to moderate maturity, moderate to low maturity, and low maturity. Responses derived from the LEAD/Self and Other reflected the perceptions of the leadership style, style range (flexibility), and style adaptability (effectiveness) of the principals. Four quadrants (scores) were developed to indicate the totals of principal dominant leadership style and supporting styles(s) as perceived by the teachers.

Meetings was developed to ascertain the teachers' and principals' perceptions of the effectiveness of school-based meetings. The perceptions of the meetings were gauged by the teachers' perceptions of how well the members worked together as a group. The questions on the Meetings questionnaire were grouped into twelve stages of problem solving: problem definition/diagnosis, solution discussion, agenda clarity/control, solution generation, orientation/summarizing, participation/resource utilization, implementation, follow-up, process analysis, solution adequacy/productivity, climate/sentiments, and decision-making resolutions. These were the criteria considered necessary for an effective meeting. The process or procedures by which the council operated were of critical importance since these meetings set the policies which guided the day-to-day operations of the school. School-based decision-making management in Kentucky is based on meetings.

Validity and Reliability

The validity of the Leader Effectiveness and Adaptability questionnaires was based on the responses of 264 managers in North America. Thirty percent of the managers were at the entry level of management, 55% were middle managers, and 14% were in the upper management level of the companies. John Greene (1980) summarized the validity and reliability of the Leader Effectiveness and Adaptability instruments as follows:

> The 12 item validation for the adaptability score ranged from .11 to .52, and 10 of the 12 coefficients (83%) were .25 or higher. Eleven coefficients were significant beyond the .01 level and one was significant at the .05 level. Each response option met the operationally defined criterion of less than 80% with respect to selection frequency.
>
> The stability of the LEAD/Self and Other was moderately strong. In two administrations across a six-week interval, 75% of the managers maintained their alternate style. The contingency coefficients were both .71 and each was significant ($p < .01$). The correlation for the adaptability scores was .69 ($p < .01$). The LEAD/Self and Other scores remained relatively stable across time, and the user may rely upon the results as consistent measures. (p. 1)

Along with Greene, the *Ninth Mental Measurements Yearbook* states:

> The responses of 264 managers, ranging in age from 21 to 64, were used to standardize the LEAD/Self and Other. The managers represented a variety of managerial levels. The concurrent validity coefficients of the 12 items ranged from .11 to .52. In another study, a significant correlation of .67 was found between the adaptability scores of the managers and the independent ratings of their supervisors.
>
> Item analyses data and reliability data were also collected on the sample of 264 managers. Each response option met the operationally defined criterion of less than 80% with respect to selection frequency. The stability of the LEAD/Self and Other was moderate. In two administrations across a six-week interval, 75% of the managers maintained their dominant style and 71% maintained their alternate styles. (Mitchell, 1985, p. 1385)

The validity of the Meetings questionnaire was based on the Cooperative Project for Educational Development (COPED). An initial field test of 150 teachers and principals were administered the questionnaire in 1970. It was then included with a package of twenty other instruments and administered to more than 3,000 adults in 21 school districts. Since then the Meetings questionnaire has been used in various other studies (Lake, 1968, 1970).

Test-retest studies yielded average item reliabilities of .60. The positive sum correlates .89 with the total score, and the negative sum correlates .90 with the total score (Lake, Miles, & Earle, 1973).

There were three separate factors in determining construct validity of the Meetings instrument. Problem-solving adequacy, commitment, and decision-making effectiveness were the three factors identified in exploring construct validity. The criterion for including an item in a factor was that it must have a .50 or better loading in at least three out of four analyses. The four studies had sample sizes of 48, 122, 491, and 625. All the participants were adults employed in school systems in the COPED study (Lake, Miles, & Earle, 1973).

The original format of the Meetings questionnaire was used, with three changes. Question 65 was added by the researchers as a positively scaled item. The final open-ended question on the original instrument was dropped. The scoring rubric was changed from a six point scale to a five point scale, with one and two being positive, four and five negative, and three being neutral. A limitation of this study was that reliability was not determined on the sample obtained.

Research Design

This study used the basic correlational research design in which two scores were obtained for each individual in the study. The first was a score on the Hersey and Blanchard Leader Effectiveness and Adaptability instrument (1988). The second score was derived from the Meetings instrument developed by Miles (1968, in Lake). Maximum scores on the LEAD/Self and Other instruments were 36, with the Meetings instrument having possible scores between a +32 and a -44. In this study, positively scored council's meetings greater than 0 were considered to be effective and any negatively scored council's meetings were considered ineffective.

Procedures

The participants used in the research were teachers and principals in 26 elementary schools randomly selected based on congressional districts in the state. Once the sites were selected, each principal was contacted and agreed to have his/her school participate. Then each school office was sent a set of materials including the two instruments (LEAD/Self and Other and Meetings) with instructions for the dissemination of materials and collection of completed forms. This data collection occurred between January and February. With the returns collected, all data were entered into a database and the Statistical Program for Social Sciences (SPSS) was used to develop the descriptive and inferential statistics.

Analysis of Data

This study used the Pearson Product Moment (r) as the major tool of analysis. In one case, Hypothesis 2, a non-independent t test was run to determine significance. Other statistical treatment of data included means, standard deviations, ranges, and percentages.

Findings

In the analysis of data, three hypotheses postulated in the study were tested.

- Ho_1 - There is no statistically significant relationship between the elementary teachers' ratings of the principals' leadership effectiveness and the teachers' ratings of the effectiveness of the school-based decision-making council meetings.
- Ho_2 - There is no statistically significant difference in the teacher council members' and the teacher non-council members' ratings of the principal's leadership effectiveness.
- Ho_3 - There is no statistically significant relationship between the teachers' ratings of principal leadership range of styles and the teachers' ratings of council meetings effectiveness.

Hypothesis 1 was rejected. It was indicated that a significant relationship existed between teachers' ratings of the principal's leadership effectiveness and teachers' ratings of the effectiveness of the school council meetings (r = .62). See Table 1.

Table 1
Relationship Between Teacher's Ratings of Perceptions of Principal's Leadership Effectiveness and School-Based Council Meetings' Effectiveness

Statistic	Value
N	320
$\sum X$	537.34
$\sum Y$	537.00
$\sum X^2$	15770.69
$\sum Y^2$	15477.00
Mean of 'X' Scores	28.28
Mean of 'Y' Scores	28.26
$\sum XY$	15446.04
Pearson's r	0.62
df	318

X = principals' leadership effectiveness
Y = meetings' effectiveness

Hypothesis 2 was also rejected. As shown in Table 2, the test of this hypothesis indicated a significant difference was found between the teacher council members' and the teacher non-council members' ratings of their principal's effectiveness (t = 2.19).

Table 2
Difference in Teacher Council Members' and Teacher Non-Council Members' Ratings of Perceptions of Principal's Leadership Effectiveness

Groups	Mean	Standard Deviation	t
Council Teachers	32.0202	3.984	2.19
Non-council Teachers	30.8333	4.229	

N - Council Teachers = 72
N - Non-council Teachers = 248
$df = 318$
$p = < .05$

Hypothesis 3 was not rejected. No significant relationship existed between the teacher's ratings of the principal's leadership range, as identified by Hersey and Blanchard, and the teachers' ratings of council meetings' effectiveness, $r = 0.0$, $df = 318$.

Conclusions and Discussion

While correlation does not indicate a causal relationship, the indication is that principals who were viewed as ineffective overall conducted council meetings that were also viewed as ineffective. Perhaps this was because teacher perceptions were hard to change or because it was difficult for ineffective individuals to adapt to new methods of operating their schools. In any case, overall principal perception was correlated to successfully viewed school-based decision-making meetings, the basic source of governance in Kentucky SBDM school sites.

The findings indicated that teachers who work most closely with the principal have the most positive view of that person's effectiveness. Again, a cause-effect relationship is not indicated, but the findings of the researchers that the teachers most closely involved with the principal in the operation of the school viewed the principals most positively was significant. This leads to the recommendation that the principal become aware of teacher perceptions and how they are impacted by proximal distance. This suggested that the principal must consciously improve his/her perception among faculty members before undertaking major changes in the school setting itself. This supported Vroom and Yetton's (as cited in Hoy & Miskel, 1993) contingency theory that variables which interacted with personal attributes of the leader resulted in leader behavior which could impact the effectiveness of the organization. This situation was especially true if the principal wished to improve his/her perception among those faculty who were only distantly knowledgeable of his/her job performance. As Pfeffer (1977) stated, leadership exists in people's minds rather than representing an object reality.

The third hypothesis dealt with Hersey and Blanchard's (1988) report that leaders who display leadership styles (scores) in two or more quadrants were more effective than leaders who tend to behave in one basic leadership style. This study found no relationship between teachers' ratings of principal leadership style and council meetings' effectiveness. However, it found no principals who were viewed as being in only a one style (score) quadrant. Since there were no one style quadrant individuals identified in this study, no conclusion could be drawn other than that *all* principals in this study were viewed as multiquadrant individuals.

It should be noted that some individual teachers did score their principals as ineffective, but the total score of the teacher group in individual schools indicated the principal's leadership was viewed as effective. The researchers felt that the criteria for this study restricted the selection of principals. If the study had looked at effective versus ineffective principals, perhaps the results would have been different.

This finding may also impact the preparation of administrators and the professional development of current principals. Such skills as are involved in development of perceptions can be identified and therefore taught principals both in pre and in-service settings. It behooves the university faculty to pay more attention to the behaviors needed in developing perceptions of competence--whatever they may be.

The researchers recommend that principals build a perception of effectiveness among staff before taking on the task of school-based decision-making or at least work to develop the staff view of his/her effectiveness over and above the work of the council. Because the two views were so closely related, the perception of the principal's overall effectiveness appears to carry over to the perception of council meeting effectiveness.

The principal must go beyond his/her normal attempts to get the staff more closely and directly involved in and knowledgeable of what he/she really does in the principal's position on a regular basis. If the principal wants more positive perceptions, he/she must sharpen his/her public relations skills since each of the three constituencies, teachers, parents, and community, can have a major impact on his/her success or failure and subsequently the success of SBDM as a governance alternative.

In mandated change situations, such as the Kentucky Education Reform Act, the repercussions of error can be enormous and long lasting. If the common denominator in effective schools, the school principal, is forced to change his/her leadership style by participating in forced school-based decision-making council settings, then some councils and principals may be set up to fail. Hersey and Blanchard (1988) stated this well when they said that the major attribute which set the successful organization apart from the unsuccessful organization was dynamic and effective leadership. While this research does not answer this question, it lends credence to the fact that this question needs to be asked and answered before SBDM schools with their concomitant impact on children are allowed to venture too far with their restructuring.

Recommendations for Further Research

The authors of this research study recommend:

1. That further study be undertaken based on identified effective and ineffective principals.
2. A replication of this study enlarging the number of principals needs to be conducted.

References

Afsahi, J. (1990). *Local school leadership council meetings: Members perceptions of their effectiveness.* Unpublished doctoral dissertation, University of LaVern, LaVern, CA.

Bahrenfuss, R. (1992). Four years later– How Greece, N.Y., uses site-based management. *Educational Leadership, 50,* 42-44.

Barth, R. (1990). A personal vision of a good school. *Phi Delta Kappan, 69,* 639-642.

Bell, T. (1983). *A nation at risk: The imperative for educational reform.* National Commission on Excellence in Education. Washington, DC: U.S. Government Printing Office.

Bergman, A. (1992). Lessons for principals from site-based management. *Educational Leadership, 50,* 48-51.

Blase, J. (1987). Dimensions of effective school leadership: The teacher's perspective. *American Educational Research Journal, 24,* 589-610.

Brandt, R. (1982). On school improvement: A conversation with Ron Edmonds. *Educational Leadership, 40,* 12-15.

Bryk, A., Easton, J., Kerbow, D., Rollow, S., & Sebring, P. (1994). The state of Chicago school reform. *Phi Delta Kappan, 76,* 74-78.

Calder, B. (1977). New directions in organizational behavior. In B. M. Straw, & G. R. Salancik, (Eds.), Chicago: St. Clair Press.

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st century.* Report of the Carnegie Task Force on Teaching as a Profession. Washington, DC: Author.

Casner-Lotto, J. (1988). Expanding the teachers' role: Hammond's school improvement process. *Phi Delta Kappan, 69,* 349-353.

Conley, S., & Bacharach, S. (1990). From school-site management to participatory school-site management. *Phi Delta Kappan, 71,* 539-544.

Cross, R. (1981). What makes an effective principal? *Principal, 60*(4), 19-22.

David, J. (1989). Synthesis of research on school-based management. *Educational Leadership, 46,* 45-53.

Dreyfuss, G. (1988). Dade county opens the door to site decisions. *The School Administrator, 45*(7), 12-13, 15.

Fiedler, F. (1967). *A theory of leadership effectiveness.* New York: McGraw Hill.

Fitch, C. (1991). School-based management councils in the United States and Chicago. *Illinois Schools Journal, 71*(1), 3-9.

Gleason, S, Donohue, N., & Leader, G. (1996). Boston revisits school-based management. *Educational Leadership, 53,* 24-27.

Gomez, J. (1989). The path to school-based management isn't smooth, but we're scaling the obstacles one-by-one. *The American School Board Journal, 176*(10), 20-22.

Greene, J. (1980). *Lead-Self Manual.* Milford, CN: University of Bridgeport.

Hammersley, M. (1990). What's wrong with ethnography? *Sociology, 24,* 597-615.

Hammond, L. (1990). Achieving our goals: Superficial or structural reforms. *Phi Delta Kappan, 72,* 286-295.

Hansen, E. (1991). Educational restructuring in the U.S.A.: Movements of the 1980's. *Journal of Educational Administration, 29*(4), 30-38.

Hersey, P., & Blanchard, K. (1988). *Leader effectiveness and adaptability description: Self and other scoring matrix and analysis.* San Diego: Leadership Studies.

Hersey, P., & Blanchard, K. (1987). *Management of organizational behavior utilizing human resources.* Englewood Cliffs, New Jersey: Prentice Hall.

The Holmes Group. (1986). *Tomorrow's teachers, a report of the Holmes group.* East Lansing, MI: Author.

Hoy, W., & Miskel, C. (1993). *Educational administrative theory, research, and practice.* (4th. ed.). New York: Random House.

Klopf, G., Scheldon, E., & Brennan, K. (1982). The essentials of effectiveness: A job description for principals. *Principal, 151,* 4, 35-38.

Koontz, H., & O'Donnell, C. (1959). *Principals of management.* (2nd ed.) New York: McGraw-Hill.

Lake, D. (Ed.). (1968). *Cooperative project for educational development. Final Report.* Washington, DC: U.S. Department of Education. (ERIC Document Reproduction Service No. ED 021 338)

Lake, D. (Ed.). (1970). *Cooperative project for educational development. Appendix: Coding manual.* Washington, DC: U.S. Department of Education. (ERIC Document Reproduction Service No. ED 042 266)

Lake, D., Miles, M., & Earle, R., Jr. (1973). *Measuring human behavior.* New York: Teacher's College Press.

Lieberman, A. (1988a). Expanding the leadership team. *Educational Leadership, 45,* 4-8.

Lieberman, A. (1988b). Teachers and principals: Turf, tension, and new tasks. *Phi Delta Kappan, 69,* 648-653.

Lunenburg, F., & Ornstein, A. (1996). *Educational administration.* Belmont, CA: Wadsworth.

Manasse, L. (1982). Effective principals: Effective at what? *Principal, 61*(4), 10-75.

McClelland, D. (1965). Toward a theory of motive acquisition. *American Psychologist, 20,* 321-333.

Meadows, B. J. (1990). The rewards and risks of shared leadership. *Phi Delta Kappan, 71*, 545-548.

Mentell, E. (1993). Implementing site-based management: Overcoming the obstacles. *NASSP Bulletin, 77*, 97-102.

Miles, M. (1968). Meetings. In D. Lake, (Ed.), *Cooperative project for educational development. Final report*. Washington, DC: U.S. Department of Education. (ERIC Document Reproduction Service No. Ed 021 338, pp. 68-81).

Mitchell, J. (Ed.). (1985). *The Ninth Mental Measurement Yearbook*. Buros Institute of Mental Measurements. Lincoln: University of Nebraska Press.

Mitchell, D., & Tucker, S. (1992). Leadership as a way of thinking. *Educational Leadership, 49*, 30-35.

Mullen, B., Symons, C., Hu, L., & Salas, E. (1989). Group site, leadership behavior, and subordinate satisfaction. *The Journal of General Psychology, 116*, 155-169.

Pascarella, S., & Lunenburg, F. (1988). A field test of Hersey and Blanchard's situational leadership theory in a school setting. *College Student Journal, 22*, 33-37.

Pfeffer, J. (1977). The ambiguity of leadership. *Academy of Management Review, 2*, 104-112.

O'Neil, J. (1996). On tapping the power of school-based management: A conversation with Michael Strembitsky. *Educational Leadership, 53*, 66-70.

Raywid, M. (1990). The evolving effort to improve schools: Pseudo-reform, incremental reform, and restructuring. *Phi Delta Kappan, 72*, 139-143.

Rist, M. (1990). Chicago decentralizes. *The American School Board Journal, 177*(9), 21-24, 36.

Rothberg, R., & Hill, M. (1992). The "Foxfire" principal: Managing the restructured school. *Journal of School Leadership, 2*, 410-418.

Sevener, D. (1991, May). Revisiting the 1985 education reforms: Is the 'old school bus' running better? *Illinois Issues*, 14-16.

Sexton, M., & Switzer, K. (1977). Educational leadership: No longer a potpourri. *Educational Leadership, 35*, 19-24.

Sirotnik, K., & Clark, R. (1988). School-centered decision-making and renewal. *Phi Delta Kappan, 69*, 660-664.

Spilman, C. (1996). Transforming an urban school. *Educational Leadership, 53*, 34-39.

Taylor, B., & Levine, D. (1991). Effective schools' projects and school-based management. *Phi Delta Kappan, 72*, 389-393.

Terry, G. (1960). *Principles of management*. (3rd. Ed.) Homewood, IL: Irwin.

Thomas, M. (1977). No perfect model: The complexities of educational leadership. *NASSP Bulletin, 61*, 34-40.

Timar, T. (1990). The politics of school restructuring. *The Education Digest*, May, 7-9.

Timar, T., & Kirp, D. (1989). Educational reform in the 1980's: Lessons from the states. *Phi Delta Kappan, 70*, 504-517.

U.S. Department of Commerce. (1990). *Census Bureau Population and Housing Characteristics in Kentucky* (p. 28). Washington, DC: U.S. Government Printing Office.

U.S. Department of Education. (1991) *America 2000: An education strategy*. Washington, DC: Author.

Vroom, V. (1976). Leadership handbook of industrial and organizational psychology. In M. Dunnette (Ed.), Chicago: Rand McNally.

Walberg, H., & Niemiec, R. (1994). Is Chicago school reform working? *Phi Delta Kappan, 75*, 713-715.

Whaley, K. (1994). Leadership and teacher job satisfaction. *NASSP Bulletin, 78*, 46-50.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 49-56

Prevalence and Identification of Attention-Deficit Hyperactivity Disorder in a Mid-Southern State

Christine E. Daley
Clinch County Schools, Homerville, GA

Harold Griffin
University of Central Arkansas

Anthony J. Onwuegbuzie
Valdosta State University

The purpose of this study was to determine the prevalence of Attention-Deficit Hyperactivity Disorder (ADHD) among school children in a mid-southern state, and to gather relevant information to assist school districts in planning appropriate educational interventions. The ADHD Survey (ADDS) was mailed to 311 school superintendents; 128 (41.1%) were returned. Findings revealed that, overall, 3% of students in the state are identified as ADHD, although in some districts, as many as 25% of students have received this diagnosis. The vast majority of school districts utilize some type of behavior rating scales/checklists in identifying children with ADHD. Ritalin is taken by ADHD students in all districts. Other medications in common use include Cylert, Dexedrine, Tofranil, Norpramin, and Adderall. The administration of medications is supervised most often by nurses/nursing personnel (45.3%). However, 32% of the districts reported that "multiple" dispensers are responsible for the delivery of prescription drugs. Behavior modification techniques are the most frequently used supplement to medication (67.9%). Medical evaluations are typically the first step in the evaluation process (52.1%), although only 64% of the districts reported using a physician's report in arriving at a diagnosis of ADHD. The implications of these findings are discussed, as well as recommendations for future research.

Attention Deficit Hyperactivity Disorder (ADHD) is probably the most widely researched and best known of any of the childhood behavioral disorders, having received significant notice in the psychological, educational, and medical literature for the past decade. Characterized primarily by inattention, impulsivity, and motor restlessness, ADHD is presumed to be the result of some underlying neurological dysfunction (Heilman, Voeller, & Nadeau, 1991; Riccio, Hynd, Cohen, & Gonzalez, 1993; Voeller, 1991) that manifests itself in the preschool years.

In addition to these fundamental difficulties, several other symptoms have been associated with ADHD, chief of which is poor academic performance. Children with ADHD are two to three times more likely than other children to be retained in grade before reaching high school (Greenberg & Horn, 1991), and up to 40% may eventually be placed in formal special education programs for children with learning disabilities or behavioral disorders (Barkley, 1990).

It also has been demonstrated that children with ADHD exhibit more language difficulties (Barkley, DuPaul, & McMurray, 1990; Hartsough & Lambert, 1985); more minor physical anomalies and health problems (Firestone, Lewy, & Douglas, 1976; Hartsough & Lambert, 1985); more sleep problems (Trommer, Hoeppner, Rosenberg, Armstrong, & Rothstein, 1988); more difficulties with problem-solving and organizational strategies (Hamlett, Pellegrini, & Conners, 1987); poorer motor coordination (Barkley et al., 1990); and a greater degree of difficulty with oppositional and defiant behavior, aggressiveness, and conduct problems (Barkley et al., 1990; Loney & Milich, 1982) than do normal children. Not surprisingly, therefore, it is estimated that more than 50% of children with ADHD also have significant difficulties in social relationships with other children (Pelham & Bender, 1982).

Despite the extensive research on this disorder, the prevalence of ADHD remains in question (Barkley,

---

Christine E. Daley is a school psychologist in the Clinch County Schools, Homerville, GA. Harold Griffin is an associate professor in the Department of Administration and Secondary Education at the University of Central Arkansas. Anthony J. Onwuegbuzie is an assistant professor in the Department of Educational Leadership at Valdosta State University. The authors of this paper wish to thank Mr. Joe Hundley and the Center for Academic Excellence at the University of Central Arkansas' College of Education, for administrative and financial support in the completion of this study. Correspondence should be addressed to: Anthony Onwuegbuzie, Department of Educational Leadership, College of Education, Valdosta State University, Valdosta, Georgia, 31698. Phone (912) 333-5924. Email: tonwuegb@valdosta.edu

1990). It is estimated that children with ADHD constitute up to one-half of the referrals to psychiatric clinics in the United States (Barkley, 1990) and represent approximately 3-9% of the school-aged population nationwide (American Psychiatric Association, 1994). Regardless, prevalence estimates have varied widely as a function of disparities in defining symptoms, instrumentation and data collection procedures, and information sources (Barkley, 1990). In addition to methodological issues, problems with differential diagnosis and comorbidity of ADHD with other disorders may also impact resulting prevalence rates (Epstein, Shaywitz, Shaywitz, & Woolston, 1991; Riccio, Gonzalez, & Hynd, 1994).

The general lack of consensus as to the best method for defining ADHD may represent the greatest barrier to obtaining accurate prevalence information. Although the disorder has been characterized as neurological in nature (Heilman et al., 1991; Riccio et al., 1993; Voeller, 1991), its diagnosis typically is based on behavioral criteria included in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV; American Psychiatric Association, 1994). These criteria, which require the presence of six of 18 behaviors, exceeding a subjectively determined level of impairment in academic, social, or occupational functioning, result in the potential for any number of different combinations which could lead to a diagnosis of ADHD (Barkley, 1990). Clearly, such marked heterogeneity in form and severity precludes the precise measurement of the extent of the problem and has even compelled some researchers (e.g., Barkley, 1982; Bloomingdale & Sergeant, 1988) to formulate their own definitions of ADHD in order to select subjects for study.

Another obstacle to accurate prevalence estimates involves the variety of instrumentation and data collection procedures utilized in arriving at an ADHD diagnosis. Despite its status as a neurologically-based disorder, there are no established biochemical markers specific to ADHD (Block, 1996). Thus, the preponderance of data are collected via interviews, behavioral observations, and rating scales, even in many cases where the diagnostic avenue has been a *medical* evaluation. Although rating scales are often portrayed as more objective than either interviews or observations, they are not without difficulty. Dykman, Ackerman, and Raney (1992) identified 42 rating scales that have been used to diagnose ADHD. Of these, according to Dykman et al., the original Conners (Conners, 1969, 1970) and the Achenbach rating scales (Achenbach, 1991) have been used more widely in studying ADHD than any others. However, at present, there are no empirical indicators on these scales that consistently identify children with ADHD (Gordon, 1991); there are no valid cutoff points which accurately identify ADHD students (Taylor, 1986); nor have these measures been revised to reflect DSM-IV criteria (Dykman et al., 1992). Although some more recently developed instruments, namely the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) and Attention Deficit Disorders Evaluation Scale (ADDES; McCarney, 1989a, 1989b) show promise, there has been insufficient research to demonstrate their diagnostic utility (Dykman et al., 1992).

In a related issue, diagnosis–and thus, prevalence–of ADHD may be impacted by the particular informants involved in the assessment process. For example, parents and/or teachers may be inaccurate in reporting children's behavior, thereby hindering reliable identification.

Finally, a major challenge in arriving at prevalence data is distinguishing ADHD from other related psychiatric syndromes. ADHD has been found to co-exist with virtually every disorder of childhood and adolescence, including mental retardation, substance abuse, Tourette's Syndrome, conduct disorder, oppositional defiant disorder, various mood and anxiety disorders, borderline personality, and learning disorders (Dykman et al., 1992). Biederman, Newcorn, and Sprich (1991), in a review of the literature on disorders frequently co-occurring with ADHD, reported that 30-50% of children with ADHD may also be diagnosed with conduct disorder; 35% with oppositional defiant disorder; 15-75% with mood disorders; and 25% with anxiety disorders. Sixty percent of children with Tourette's Syndrome and 25% of those with borderline personality have a comorbid attention deficit disorder. In addition, ADHD occurs three to four times more frequently in mentally retarded children than in normals, particularly in the mildly retarded group (Biederman et al., 1991). Learning disabilities (LD) also are prevalent among children with ADHD. Ackerman and Dykman (1990) suggest that approximately one-third to one-half of all ADHD children are LD, depending on the criteria one uses to label a child LD. Among LD populations, the reported prevalence of ADHD has varied from 48% (Holborow & Berry, 1986) to 80% (Safer & Allen, 1976). This considerable overlap with a number of other disorders not only raises the question of ADHD's validity as a distinct diagnostic entity but also has the potential to impact significantly upon reported prevalence rates.

The behavioral heterogeneity and high levels of comorbidity characteristically associated with ADHD also have important implications for the differential effectiveness of various treatment approaches, particularly pharmacological ones. Some have argued (e.g., Block, 1996) that once a diagnosis of ADHD is made, physicians all too frequently move on to prescribing stimulant medications, such as Ritalin, Cylert, or Dexedrine. Despite the fact that these drugs cannot do all things for such a heterogeneous group, the use of medication in the treatment of children with ADHD is widely accepted and commonly practiced

(Barkley, 1990; Greenhill, 1992). Indeed, Reid, Maag, Vasa, and Wright (1994) reported that 90% of their ADHD sample was receiving medication; Wolraich et al. (1990) reported a medication rate of 88%.

This widespread use of drug therapy also bears a considerable impact on educational systems. Because children spend a significant proportion of their day in school, medication use among students with ADHD often places teachers, school nurses, and administrators in the role of medical managers. This function carries with it a number of responsibilities, including accountability for controlled substances, preservation of a child's right to confidentiality, monitoring of medication efficacy, and awareness of possible side effects (Reid et al., 1994). The potential for student risk is high when schools are not mindful of these obligations.

Unfortunately, we have few specifics about how and how well students with ADHD are being served in schools because of a dearth of literature on this topic (Chesapeake Institute, 1992; Reid, Maag, & Vasa, 1993; Reid et al., 1994). Placements for these students may range from regular classroom with no services, to regular classroom with accommodations, to a variety of special education settings. In any case, as an adjunct to pharmacological therapy, they likely experience an assortment of non-medical treatments, including positive reinforcement, token economies, contingency contracting, response cost, and time out (DuPaul, Guevremont, & Barkley, 1991; Franks, 1987; Wallander & Hubert, 1985). Overall, behavior therapy appears to have fared well in the schools. Gadow (1985), in a review of 16 studies comparing the use of medication and behavioral interventions, concluded that the latter was far more effective in remediating academic difficulties. Additionally, such interventions, when coordinated with parent involvement, are believed by many to facilitate the generalization of treatment effects across settings and behavioral domains (Barkley, 1990).

Because children with ADHD have specific needs that must be met in order for them to achieve academic success, it is imperative that school systems recognize these students early and develop appropriate educational programming. Epidemiological research can assist this planning by providing a best estimate of the prevalence of the disorder within a given population (Francis, 1993). Unfortunately, since few studies have examined ADHD among school-based samples (Chesapeake Institute, 1992; Reid et al., 1993, 1994), we know little about the methods used to identify ADHD students, the types of placements and services they are obtaining, or the treatments and interventions they are receiving. Thus, the primary purpose of the present investigation was to estimate the prevalence of ADHD among school children in a midsouthern state and to gather relevant information to assist school districts in planning appropriate educational interventions.

Method

Instruments

The ADHD Survey (ADDS), which was developed specifically for this study, contained 11 items and was divided into five major areas of concern, namely: (1) prevalence of ADHD, (2) diagnosis of ADHD, (3) placement of ADHD students, (4) interventions for ADHD, and (5) referral process. Content validity of the ADDS was established via review by a panel of experts. Modifications were made as recommended. A summary of the final version of the survey is found in Table 1. In order to ensure the anonymity of respondents, and thus to increase the validity of their responses, the questionnaires were not coded.

Subjects and Procedure

The ADDS was mailed to all 311 superintendents of school districts in the state in which the study was undertaken. The superintendents were given three weeks in which to respond. This secured an initial response rate of approximately 30%. When the three weeks had elapsed, a second mailout to all superintendents was undertaken, increasing the response rate by an additional 10%. Thus, overall, 128 superintendents (41.1%) returned the ADDS, representing school districts with enrollments ranging from 90 to 20,328 students ($M = 1671.6$, $SD = 2634.3$).

Results

Prevalence of ADHD

The number of children in each school district identified as ADHD ranged from 1 to 563 students ($M = 43.6$, $SD = 92.7$). As such, the ADHD prevalence rate ranged from 0.21% to 25.02% per school district, with an overall mean of 3.03% ($SD = 3.37\%$). Frequency distributions of the prevalence rates are reported in Table 2. All school districts identified ADHD students in the elementary and middle school grades ($M = 2.8$, $SD = 2.4$). Indeed, 60.0% of school districts identified ADHD children by Grade 1, with 78.1% rendering a diagnosis by Grade 5, and 92.4% by Grade 6.

Diagnosis of ADHD

With regard to instrumentation utilized in diagnosing ADHD, one-third of school districts reported using only the ADDES. The next most commonly utilized (28.8%) method involved a battery of tests, including tests of intelligence, achievement, personality, motor skills, and perceptual skills, as well as behavior rating scales/checklists. This was followed by behavior rating scales/checklists only (17.1%); and ADDES and behavior rating scales/checklists only (3.6%). As many as 13.5% did not use any instruments in identifying ADHD children.

Aside from diagnostic instruments, a physician's diagnosis/report (64.1%) was cited as the most common criterion utilized in making a determination of ADHD. This was followed, respectively, by teacher(s)' observation/report (50.8%), parent(s)' observation/report (30.5%), report/diagnosis by Child Study Centers/other agencies (8.6%), school psychology specialists' observation/report (6.3%), committee decisions (6.3%), student achievement (4.7%), and school and/or discipline records (3.1%).

Table 1
Summary of Survey Items

Area of Inquiry	Specific Items
Prevalence of ADHD	How many students enrolled in your school district are currently diagnosed ADHD? At what grade level is a diagnosis of ADHD typically made in your school district?
Diagnosis of ADHD	What diagnostic test(s) are currently being used in your district to identify ADHD children? Aside from diagnostic instruments, what other criteria are utilized in making a determination of ADHD?
Placement of ADHD Students	What percentage of ADHD students in your school district are served under IDEA and Section 504? What percentage of ADHD students in your school district are not served?
Interventions for ADHD	What medications are being taken by your students with ADHD? Do school personnel supervise the administration of medication to your ADHD students? If so, who? Aside from medication, what other interventions for ADHD are being utilized in your school district? Who in your school district is responsible for overseeing the design/implementation/follow-through on accommodations made for ADHD students being served under Section 504? Does your district utilize the concept of multidisciplinary child study teams/student assistance teams/intervention teams?
Referral Process Utilized	Describe the referral source utilized in your school district for identifying ADHD children. Who typically is the referral source? To whom is the child initially referred? Which is conducted first--a medical evaluation or a complete psychoeducational evaluation? What role (if any) do the multidisciplinary child study teams/student assistance teams/intervention teams serve in the referral process?

Table 2
Frequency Distribution of Prevalence Rates

Prevalence Rates*	Percentage of School Districts
0.0 - <1.0	12.5
1.0 - <2.0	28.9
2.0 - <3.0	22.6
3.0 - <4.0	12.5
4.0 - <5.0	7.8
5.0 - <6.0	3.1
6.0 - <7.0	1.6
7.0 - <8.0	1.6
8.0 - <9.0	0.8
9.0 - <10.0	0.0
≥10.0	3.9

25th percentile = 1.3%; median = 2.2%; 75th percentile = 3.7%; semi-interquartile range = 1.2%

*4.7% of school districts did not report prevalence rates

Placement of ADHD Students

On average, 39.1% (*SD* = 32.5%) of ADHD students in each school district are served under the Individuals with Disabilities Education Act (IDEA), 19.6% (*SD* = 27.6%) are served under Section 504 of the Rehabilitation Act of 1973, and 40.1% (*SD* = 36.6%) receive no services.

Interventions for ADHD

With respect to medications administered to ADHD students, Ritalin was reported as the most common--taken by ADHD students in all school districts. The administration of Cylert (53.5%) was the next most frequently reported, followed by Dexedrine (47.2%), Tofranil (22.0%), Norpramin (10.2%), and Adderall (7.9%). Between 0.8% and 1.6% of school districts reported use of one or more of the following by students with ADHD:

Tegretol, Thorazine, Depakene, Mellaril, Desoxyn, Prozac, Adapin/Sinequan, and a combination of vitamins.

The administration of medications is supervised most often by nurses/nursing personnel (45.3%). In addition, 7.8% of the districts reported the administration of medication by teachers, 7.0% by principals/administrative staff, and 5.5% by secretaries. Thirty-two percent of the districts reported that multiple dispensers are responsible for the administration of medications.

Aside from medications, behavior modification is the most frequently utilized intervention (67.9%). Examples of this included time out, loss of privileges, positive reinforcement, and punishment. Other interventions cited by superintendents were the use of structured classrooms (33.6%), shortened/modified assignments and/or tests (21.1%), home-school contracting (14.1%), counseling (8.6%), special seating arrangements (8.6%), change of placement/special education (7.8%), contracts (5.5%), special materials (3.9%), tutoring (3.1%), essential skills training (2.3%), staggering low/high interest materials (1.6%), alternative discipline (1.6%), social skills training (1.6%), brief activity periods (0.8%), mentoring with teachers (0.8%), diet control (0.8%), parent contract (0.8%), and textbooks on tape (0.8%). Only one superintendent reported that her/his school district utilized no interventions for ADHD other than medication.

Nearly all (92.9%) school districts had a designated Section 504 "coordinator" who was responsible for overseeing the design/implementation/follow-through on accommodations made for ADHD students. In these school districts, the individuals responsible for coordinating this provision included directors of special services (26.6%), principals (22.3%), counselors (17.0%), assistant superintendents (10.6%), superintendents (8.5%), federal program directors (6.4%), local education authority directors (4.3%), resource teachers (3.2%), and assistant principals (1.1%).

With respect to the referral process utilized in school districts for identifying ADHD children, teacher-parent combinations are the most common referral source (56.6%), followed by teachers alone (23.8%), and parents alone (14.8%). The survey revealed that ADHD children are most often referred to resource teachers (26.9%), followed by principals (21.0%), counselors (12.6%), physicians/nurses (12.6%), special committees (e.g., Section 504 personnel), a combination of regular classroom teachers, special education teachers, and principals (4.2%), a combination of counselor and principal (4.2%), regular classroom teachers (2.5%), and a combination of regular classroom teachers and counselors (1.7%). In slightly more than one-half (52.1%) of school districts, a medical evaluation preceded a psychoeducational evaluation in identifying ADHD children. In 31.4% of school districts, the reverse is true (i.e., a psychoeducational evaluation preceding a medical evaluation). The remainder of school districts either rely on the recommendation of teams (14.9%) or use both psychoeducational evaluation and medical evaluations concurrently (1.7%). Overall, 45.8% of school districts utilize the concept of multidisciplinary child study teams/student assistance teams/student intervention teams as part of the referral process. The role of these teams includes the following: to make recommendations for/against evaluation (27.3%), to collaborate on ideas for intervention (27.3%), to conduct screening/evaluations (7.2%), to coordinate the entire referral process (7.2%), to discuss progress and needs (5.5%), to make educational decisions for students suspected of having ADHD (5.5%), and to provide support for parents (5.5%).

Discussion

The ADDS revealed that approximately 3% of school-aged students in this mid-southern state are diagnosed with ADHD. This finding is consistent with the national estimate of 3-9% (APA, 1994). A somewhat disturbing finding was the fact that, in some school districts, as many as 25% of students are identified as being ADHD. This raises the possibility that inappropriate numbers of children are receiving this diagnosis. As noted earlier (Barkley, 1990), prevalence estimates may be impacted by a number of factors, including diagnostic procedures, instrumentation, and informants. It is clear from this study's findings that there is little statewide standardization in procedures for identifying ADHD children–a conclusion consistent with what appears to be a troubling national trend (Reid et al., 1993). Not only does the referral process in this state vary widely from district to district, but only 64% of the local education authorities (LEAs) report considering a physician's diagnosis in making an ADHD determination. Additionally, most of the districts appear to rely heavily on the use of a variety of behavior rating scales and checklists in arriving at a diagnosis, despite the questionable reliability and validity of these instruments noted earlier (Dykman et al., 1992; Gordon, 1991; Taylor, 1986).

With regard to the placement of children with ADHD, approximately 39% are receiving special education services under IDEA. Although the prevalence of children with ADHD who reqnire special education has not been studied directly, estimates suggest that approximately 50% are, in fact, in need of such services (Council for Exceptional Children, 1992), either because of the direct results of their attentional difficulties or because of some concomitant educational disability. This estimate would seem reasonable, given the multitude of data noted earlier (e.g., Biederman et al., 1991; Dykman et al, 1992) linking ADHD with virtually every childhood disorder. In any case, it would appear that at least some ADHD students in this state are not receiving necessary special education services.

As for the 20% of ADHD students who are being served under Section 504, although nearly all districts designated a "504 coordinator," a significant proportion (50%) of these individuals fill roles which seem considerably removed from the site of service implementation (e.g., superintendents, assistant superintendents, directors of special services, LEA directors). This raises questions of appropriate monitoring of and accountability for individual accommodation plans, which all too often, may be perused and forgotten by the overburdened regular classroom teacher. Indeed, there is some evidence to suggest that teachers in the general classroom feel unprepared to deal with the needs of ADHD students (Reid et al., 1994).

Although this survey did not ask respondents to estimate the number of ADHD children receiving pharmacological treatment, it is clear that a spectrum of stimulant, antidepressant, antiseizure, and anti-hypertensive medications are being used by children in every district. Perhaps of gravest concern in this study was the finding of the variety of individuals responsible for the administration of these controlled substances, with 32% of the districts reporting the use of "multiple" dispensers. As noted earlier (Reid et al., 1994), the culpability inherent in medical management is considerable, not to mention the risk to students in situations in which teachers, administrators, and others may be unaware of potential adverse side effects.

An encouraging finding in this study was the report of extensive usage of non-pharmacological intervention as a corollary to drug therapy for ADHD students. Indeed, only one school district indicated that it used no additional treatment methods. Given the reported superiority of behavioral strategies in the management of ADHD (Gadow, 1985), this is clearly representative of a best practices approach.

Implications and Recommendations

It is apparent that differences in conceptualization and diagnostic procedures are major factors in the estimation of prevalence rates for ADHD. A priority of research and practice must be, therefore, a consensus regarding the defining features of this disorder and a standardization of approaches to identification and differential diagnosis.

Future research also should examine the specific disability conditions which qualify some ADHD students for special education placement, comparing the characteristics of these students to those who are maintained in the regular classroom.

School districts must establish a foolproof system of follow-up and accountability for the implementation and evaluation of individual accommodations plans written for ADHD students who are being served under Section 504. Regular classroom teachers must be equipped with knowledge of ADHD and an arsenal of skills to handle the difficulties experienced by these students in the inclusive environment.

Teachers, administrators, and staff who are involved in dispensing medication to students with ADHD should be educated in potential adverse reactions and side effects. Schools should maintain a reliable line of communication with parents and physicians in the event that any problems related to medication arise.

Finally, given the paucity of research on ADHD in the schools, future investigations should focus on accumulating data in the academic environment, where the disorder is, arguably, most pernicious.

It is imperative that we design and implement appropriate interventions to ensure that children with ADHD experience success in school and beyond. Only with additional knowledge and understanding of this disorder will we have the tools to accomplish this goal.

References

Achenbach, T. M. (1991). *Integrative guide for the 1991 CBCL/4-18, YSR, & TRF Profile*. Burlington, VT: University of Vermont Department of Psychiatry.

Ackerman, P. T., & Dykman, R. A. (1990). Prevalence of additional diagnoses in ADD and learning disabled children. *Advances in Learning and Behavioral Disabilities, 6*, 1-25.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barkley, R. A. (1982). In B. B. Lahey & A. E. Kazdin (Eds.), Specific guidelines for defining hyperactivity in children. *Advances in clinical child psychology* (Vol. 5, pp. 137-180). New York: Plenum.

Barkley, R. A. (1990). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment*. New York: Guilford.

Barkley, R. A., DuPaul, G. J., & McMurray, M. B. (1990). A comprehensive evaluation of Attention Deficit Disorder with and without Hyperactivity defined by research criteria. *Journal of Consulting and Clinical Psychology, 58*, 775-789.

Biederman, J., Newcorn, J., & Sprich, S. (1991). Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. *American Journal of Psychiatry, 148*, 564-577.

Block, M. A. (1996). *No more Ritalin: Treating ADHD without drugs.* Toronto: Kensington.

Bloomingdale, L. M., & Sergeant, J. (1988). *Attention deficit disorder: Criteria, cognition, intervention.* New York: Pergamon.

Chesapeake Institute (1992). *Executive summaries of research syntheses and promising practices on the education of children with attention deficit disorder.* Washington, DC: Author. (ERIC Document Reproduction Service No. ED 363 083)

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry, 126,* 884-888.

Conners, C. K. (1970). Symptom patterns in hyperkinetic, neurotic, and normal children. *Child Development, 41,* 667-682.

Council for Exceptional Children (1992). *Children with ADD: A shared responsibility.* Reston, VA: Author. (ERIC Document Reproduction Service No. ED 349 716)

DuPaul, G. J., Guevremont, D. C., & Barkley, R. A. (1991). Attention-deficit hyperactivity disorder. In T. R. Kratochwill & R. J. Morris (Eds.), *The practice of child therapy* (2nd ed., pp. 115-144). New York: Pergamon.

Dykman, R. A., Ackerman, P. T., & Raney, T. J. (1992). *Assessment and characteristics of children with attention deficit disorder.* Little Rock, AR: Department of Pediatrics, Arkansas Children's Hospital. (ERIC Document Reproduction Service No. ED 363 084)

Epstein, M. A., Shaywitz, S. E., Shaywitz, B. A., & Woolston, J. L. (1991). The boundaries of attention deficit disorder. *Journal of Learning Disabilities, 24,* 78-85.

Firestone, P. Lewy, F., & Douglas, V. I. (1976). Hyperactivity and physical anomalies. *Canadian Psychiatric Association Journal, 21,* 23-26.

Francis, G. (1993). A Prevalence study: ADHD in elementary school children. *Canadian Journal of School Psychology, 9,* 16-27.

Franks, C. M. (1987). Behavior therapy with children and adolescents. In G. T. Wilson, C.M., Franks, P.C. Kendall, & J.P. Foreyt (Eds.), *Review of behavior therapy: Theory and practice* (Vol. 2, pp. 234-287). New York: Guilford.

Gadow, K. D. (1985). Relative efficacy of pharmacological, behavioral, and combination treatments for enhancing academic performance. *Clinical Psychology Review, 5,* 513-533.

Gordon, M. (1991). *ADHD/hyperactivity: A consumer's guide.* Dewitt, NY: GSI.

Greenberg, G. S., & Horn, W. F. (1991). *Attention deficit hyperactivity disorder: Questions and answers for parents.* Champaign, IL: Research Press.

Greenhill, L. L. (1992). Pharmacological treatment of attention deficit hyperactivity disorder. *Psychiatric Clinics of North America, 15*(1), 1-26.

Hamlett, K. W., Pellegrini, D. S. & Conners, C. K. (1987). An investigation of executive processes in the problem-solving of attention deficit disorder-hyperactive children. *Journal of Pediatric Psychology, 12,* 227-240.

Hartsough, C. S., & Lambert, N. M. (1985). Medical factors in hyperactive and normal children: Prenatal, developmental, and health history findings. *American Journal of Orthopsychiatry, 55,* 190-210.

Heilman, K. M., Voeller, K. K. S., & Nadeau, S. E. (1991). A possible pathophysiological substrate of attention deficit hyperactivity disorder. *Journal of Child Neurology, 6* (Suppl.), S76-S81.

Holborow, P. L., & Berry, P. S. (1986). Hyperactivity and learning difficulties. *Journal of Learning Disabilities, 19,* 426-431.

Loney, J., & Milich, R. (1982). Hyperactivity, inattention, and aggression in clinical practice. In D. Routh & M. Wolraich (Eds.), *Advances in developmental and behavioral pediatrics* (Vol. 3, pp. 113-147). Greenwich, CT: JAI Press.

McCarney, S. B. (1989a). *The Attention Deficit Disorders Evaluation Scale - Home Version technical manual.* Columbia, MI: Hawthorne Educational Services.

McCarney, S. B. (1989b). *The Attention Deficit Disorders Evaluation Scale - School Version technical manual.* Columbia, MI: Hawthorne Educational Services.

Pelham, W. E., & Bender, M. E. (1982). Peer relationships in hyperactive children: Description and treatment. In K. D. Gadow & I. Bialer (Eds.), *Advances in learning and behavioral disabilities* (Vol. 1, pp. 365-436). Greenwich, CT: JAI Press.

Reid, R., Maag, J. W., & Vasa, S. F. (1993). Attention deficit hyperactivity disorder as a disability category: A critique. *Exceptional Children, 60,* 198-214.

Reid, R., Maag, J. W., Vasa, S. F., & Wright, G. (1994). Who are the children with attention deficit-hyperactivity disorder? A school-based survey. *Journal of Special Education, 28,* 117-137.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children.* Circle Pines, MN: American Guidance Service.

Riccio, C.A., Gonzalez, J. J., & Hynd, G. W. (1994). Attention-deficit hyperactivity disorder (ADHD) and learning disabilities. *Learning Disability Quarterly, 17,* 311-322.

Riccio, C. A. Hynd, G. W., Cohen, M. J., & Gonzalez, J. J. (1993). Neurological basis of attention-deficit hyperactivity disorder. *Exceptional Children, 60*, 118-124.

Safer, D. J., & Allen, R. D. (1976). *Hyperactive children: Diagnosis and management*. Baltimore: University Park Press.

Taylor, E. (1986). *The overactive child*. Clinics in Developmental Medicine No. 97. Philadelphia: J. B. Lippincott.

Trommer, B. L., Hoeppner, J. B., Rosenberg, R. S., Armstrong, K. J., & Rothstein, J. A. (1988). Sleep disturbances in children with attention deficit disorder. *Annals of Neurology, 24*, 325.

Voeller, K. K. S. (1991). Toward a neurobiologic nosology of attention deficit hyperactivity disorder. *Journal of Child Neurology, 6* (Suppl.), S2-S8.

Wallander, J. L., & Hubert, N. C. (1985). Long-term prognosis for children with attention deficit disorder with hyperactivity (ADD/H). In B.B. Lahey & A.E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 8, pp. 114-147). New York: Plenum.

Wolraich, M.L., Lindgren, S., Stromquist, A. Milich, R., Davis, C., & Watson, D. (1990). Stimulant medication use by primary care physicians in the treatment of attention deficit hyperactivity disorder. *Pediatrics, 86*, 95-101.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 57-63

A Case Study of an In-School Suspension Program in a Rural High School Setting

Tammye Turpin
Louisiana Educational Consortium

Dawn T. Hardin
Northeast Louisiana University

Many schools implement new in-school suspension (ISS) programs to comply with alternative education program requirements. Rural systems face mandated program requirements with little resource allocation. One solution to the expense of an ISS instructor is the use of a camera and television to monitor ISS student behavior. This study investigated a rural high school's first year ISS program utilizing a camera and television to monitor students. A triangulated evaluation design was used to collect data concerning administrators', teachers' and students' perceptions of the new program. Findings indicated that teachers and students perceived the new ISS program as positively impacting discipline. Areas perceived in need of improvement included the targeting of students' sleeping behavior and the development of strategies to reduce the number of students repeatedly involved in ISS.

Introduction

An in-school suspension program was initiated in a rural high school for the 1995-96 school year. A primary goal of in-school suspension (ISS) programs is to provide a positive alternative to out-of-school suspension, yet remove disruptive students from classes in which they cause major discipline problems (Montgomery County Public Schools, 1981). Another goal of the program is to provide disruptive students removed from classes an opportunity to complete class assignments. Although research findings vary on the effect of ISS and academic performance, requiring these students to complete class work at school instead of sending them home may avert academic failure which is common for these students (Silvey, 1995; Siskind, 1993).

This study used an observational case study approach to gather data on a new ISS program instituted in a rural high school setting and to examine the effectiveness of ISS for improved discipline. Teachers and administrators were asked to discuss perceived effectiveness of the ISS program through structured interviews containing open-ended questions. A majority of students at the school completed open-ended questionnaires describing their perceptions of the ISS program and the impact of the program on their behavior. The ISS room was also observed by one of the researchers to gather information on students' behavior while in ISS.

Review of Related Literature

The use of out-of-school suspension is an unsatisfactory punishment for discipline problems in many situations. Out-of-school suspension in many instances results in unsupervised time off for students, especially when students come from single parent homes or have parents who work. For those students who view out-of-school suspension as a day off from school, sending them home may reinforce the behaviors an administrator is trying to correct (Jones, 1983).

An in-school suspension (ISS) program provides an alternative to some of the concerns created by out-of-school suspension. Keeping students at school with an ISS program prevents students from wandering through communities unsupervised. It provides a level of instruction for the suspended students and fosters effective communication and improved public relations with the parents of disruptive students. Furthermore, an ISS program can provide opportunities for effective counseling of troubled students which is rarely addressed in typical suspension scenarios (Disciullo, 1984).

The ISS program should not be viewed as a replacement for out-of-school suspension. Billings and Enger (1995) note that although its primary objective is to

---

Tammye Turpin is presently with the Louisiana Systemic Initiatives Program at Louisiana State University-Shreveport. Dawn Hardin serves as an Assistant Professor of Educational Leadership at Northeast Louisiana University in Monroe. Correspondence regarding this article should be directed to Dawn Hardin, Department of Educational Leadership and Counseling, 306 Strauss Hall, Northeast Louisiana University, Monroe, LA 71209 or by e-mail at edhardin@alpha.nlu.edu.

reduce the number of out-of-school suspensions, research indicates that ISS is not considered by teachers, students, parents or the community as appropriate for severe discipline problems. An ISS program can help detect problem students, hopefully helping to change their problem behavior before that behavior becomes severe. According to Fischel (1986), another primary objective of a successful ISS program is to formulate strategies to improve the behavior of disruptive students.

Factors to be addressed during the planning and implementation of a successful ISS program fall into two major areas of concern. Initially, program administration considerations, such as administrative leadership, faculty involvement and efficient communication, are essential for program success (Whitfield & Bulach, 1996). The second major area of consideration concerns the development of student routines to take place in the ISS program. These routines include the structure to be followed by students in ISS, the manner in which they will complete their assignments, and the program of counseling to address their behavioral problems.

To enhance teacher involvement and support of ISS, inservice training for faculty involved in the program, and accurate and frequent communication are essential (Siskind & Others, 1993). Teachers who do not receive accurate information regarding ISS student placement may not be aware of the student's location and may fail to provide the required class work for the student. Teachers who are informed as to why students are in ISS are much more likely to reinforce newly learned appropriate class behavior demonstrated by problem students. Teachers tend to be more supportive of ISS and become more involved in the program if information is efficiently communicated between all those involved (Corbett, 1981; Whitfield & Bulach, 1996).

Inservice for all involved in the ISS program is vital to ensure that desired goals and objectives are clearly understood. By sharing program goals and objectives, teachers will be more likely to work with ISS personnel to facilitate the proper functioning of the program (Corbett, 1981). A faculty adequately oriented to the ISS program tends to develop stronger commitments to the program philosophy, as well as a deeper understanding of the operations of the program (Sullivan, 1989).

In addition to ISS training, the assignment of a full-time qualified staff person dedicated to the ISS program is another frequently cited component of successful ISS programs. This staff person brings consistency in the administration of student guidelines to be followed in the ISS room. The ISS staff person facilitates accurate record keeping systems to assure thoroughness in communication and data collection (Sullivan, 1989).

Three components specifically related to ISS student routine consistently appear in effective ISS programs. These components include isolation, instruction, and counseling of students in ISS (Corbett, 1990; Disciullo, 1984; Jones, 1983; Mendez & Sanders, 1981; Sullivan, 1989; Weiss, 1983).

Isolation of disruptive students is necessary. Students should be completely isolated from other students with little opportunity for peer interaction while in ISS. Students should have limited lavatory usage, usually twice a day. Lunch should also take place in isolation. This is usually accomplished by bringing lunches to the students, allowing students to use the cafeteria at a specific period, or having students bring sack lunches (Disciullo, 1984; Fischel, 1986; Jones, 1983).

For an ISS program to be successful, instruction cannot be neglected. Regular class assignments must be provided by students' teachers and ISS students should be required to complete this work while they are in the program (Disciullo, 1984).

It has also been suggested that an ISS program must compensate for instruction lost in the regular classroom. Tutorial assistance should be provided and students must not be penalized for work successfully completed (Sullivan, 1989).

Counseling serves a vital, yet often overlooked role in successful ISS programs. Counseling provides students with opportunities to identify problem behavior, recognize consequences of behavior, and establish new goals of appropriate class conduct. Counseling helps students to develop better self images and improved decision making strategies, especially in the area of behavior (Sullivan, 1989).

Another vital, yet often overlooked component involves individual student follow-up procedures monitored through documentation by parents, teachers, and students. Monitoring of student progress assesses student behavior and provides the information necessary to determine the effectiveness of the ISS program (Sullivan, 1989).

To assess and evaluate the effectiveness of an ISS program, one can measure responses to four questions:

1. Does the program significantly decrease the number of instructional days lost by students involved in discipline problems?
2. Does the program significantly decrease the number of ISS incidences taking place at your school?
3. Does the program significantly decrease the number of students not following school rules?
4. Does the program significantly decrease the frequency of repeated assignment to ISS? (Weiss, 1983)

Design and Methods

A qualitative case study was conducted to obtain information and determine the effectiveness of a new in-school suspension (ISS) program initiated at a small rural high school in North Central Louisiana. This high school is the largest of nine schools in a rural parish serving 3011 students. The school is one of two in the parish serving grades nine through twelve. Two elementary schools, two middle schools and three kindergarten through twelfth grade schools serve the remaining parish students. Student enrollment at the case study school included 189 females and 175 males with a racial makeup of 57% black students, 42% white students, and 1% students of other ethnicity. The student enrollment included 32% living below the poverty level and 18.7% living in single parent households. In addition, the school reported a teen pregnancy rate of 19.9% and a drop-out rate of 7.2%.

Three areas were examined during the case study to determine the effectiveness of the school's ISS program:

1. Teacher perceptions of the ISS program's effectiveness on overall school discipline and the impact of the ISS program on classroom discipline,
2. Student perceptions of the ISS program's effectiveness on overall school discipline and the impact of the program on student behavior, and
3. Student behavior taking place in the ISS room.

Data Collection

A triangulated evaluation design was used to collect data. The goal of data collection was to determine participants' (administrators, teachers, and students) perceptions of the new ISS program. Data collection methods used included: structured interviews of administrators and teachers, open-ended questionnaires completed by students, and direct observation of students in the ISS room.

Structured Interviews. The researcher conducted individual 20 minute interviews with the administrators and teachers. The principal and assistant principal were interviewed to determine overall program operation and implementation while structured interviews of teachers were conducted to determine teachers' perceptions of the ISS program. The researcher interviewed 13 of 19, or 68%, of the full-time teachers at the school. Seventy-one percent of all faculty involved in the ISS program were interviewed during the case study. One researcher is a teacher at the school and therefore was excluded from the study. To address subjectivity and researcher bias, considerable care was exercised in determining the interview questions and procedures.

Student Questionnaires. Open-ended student questionnaires were developed by the researcher to determine students' perceptions of the ISS program. Students were given the questionnaires to complete during their English classes. Two English classes did not participate in the survey. Questionnaires were completed by 233 students in the participating classes, producing a sample of 64% of the student population.

Observations. The researcher made observations of the ISS room over a four-day period. The ISS program at this school is unique in that a television camera and monitor are used to observe student behavior. The television monitor was used by the researcher to observe the behavior of students in the ISS room at different times of the day for short periods of time (two to five minutes). Fifteen observations of student behavior took place during the four day observation period. Field notes were compiled on the 75 minutes of student behavior observed during the four day period.

Data Analysis

Data from structured interviews, open-ended student questionnaires, and observation fieldnotes were analyzed using qualitative inductive analysis. This analysis method used repeated examination of data to determine emerging categories of data. Triangulation of multiple sources was used to establish construct validity and reliability of the case study (Yin, 1994).

Findings

Teachers' Perceptions

All the interviewed teachers perceived some benefit or positive outcome of the school's ISS program. The following statements are examples of teachers' responses to the question of their general feelings or views on the ISS program:

"I love it, it is doing a great job, our discipline is much more effective this year" (Interview, 1-24, #8).

"I think it's great, it gives students a place to work, they have to do their class work instead of going home to goof off" (Interview, 1-23, #7).

"I think it's good, I'm glad we are doing it" (Interview, 1/23, #5).

"I use it and depend on it to help me teach my class" (Interview, 1/25, #11).

Another area of complete agreement among the teachers was that discipline this year at the school was better than last year. These comments included:

"This year is better, they will do more now when you send a student to the office" (Interview, 1/18, #1).

"The school I was at last year did not have an effective ISS program, this program is better" (Interview 1/25, #12).

"ISS has been an effective discipline program, because the kids hate going there" (Interview, 1/23, #5).

"I'm a little saner, less crazy, in terms of the kids than I was last year" (Interview, 1/18, #2).

Most teachers viewed the ISS program as an effective alternative to out-of-school suspension that could provide opportunities for disruptive students to pass their classes.

Four of thirteen teachers interviewed expressed that they really were not familiar with the purpose of the school's ISS program because that purpose had not been communicated to them by the administration. These comments included:

"It eliminates expulsion/suspension or a least cuts down on the number of out of school suspensions" (Interview, 1/23, #6).

"ISS allows us to hang on to these students, punish them, but still give them the chance to pass" (Interview, 1/26, #13).

"Once a student has failed your class, they have no incentive for work or good behavior, they will be even more disruptive" (Interview, 1/26, #13).

"I see it as an alternative to out-of-school suspension, paddling, and teachers having to deal with so much discipline" (Interview, 1/23, #5).

Teachers identified several characteristics that they believed contributed to the program's success. Teachers identified isolation as highly effective for stopping disruptive behavior. Isolation allowed students the time to think about the inappropriate behavior and to associate the inappropriate behavior and the resulting consequences. The completion of school work was considered another strength of the program. Students could keep up with their class work during their punishment. Another strength of the program was the atmosphere in the ISS room. Teachers perceived that students did not like the ISS room. Comments included:

"Students are away from friends, which is something that is very distasteful for students, they want to be with those friends. It is one of the most important things to a teenager" (Interview, 1/25, #12).

"Students aren't allowed to have relations with other students, it is effective because it isolates students" (Interview, 1/26, #13).

"It is working here, it was not working at my school last year because students were not isolated" (Interview, 1/25, #12).

"It also gives them (the students) time to think about what they've done that is wrong. It links misbehavior to punishment immediately" (Interview, 1/23, #6).

Of the 13 teachers interviewed, 10 stated that the major problem with the school's ISS program was that many students slept while in the ISS room. Two teachers stated that the ISS punishment was not severe enough for second and third offenses. These two teachers felt little was being done to prevent repeat ISS experiences. One teacher indicated the lack of guidelines as a major concern. Without established program guidelines, students may receive different levels of punishment for the same offense. Comments included:

"They (administration) need to make sure students are not sleeping" (Interview, 1/23, #5).

"I don't like to go by (the television monitor) and see them with their heads on their desks sleeping. We need to make sure they do work the whole time" (Interview, 1/18, #2).

"I've never seen any student get more than one day no matter how many times sent. It doesn't seem severe enough, them only getting one day" (Interview, 1/18, #1).

"The principal needs to keep up with how many times a student has been in the office and increase the severity of the offense when it is their second or third time" (Interview, 1/18, #1).

Of the 13 teachers interviewed, 11 stated the ISS program could be improved by hiring someone to oversee the program and stay in the room with the students at all times. The problem of students sleeping during ISS could be overcome by having a staff person in the room to require students to work at all times. Another recommendation for improvement suggested parent notification. Three teachers stated that there lacked enough parental involvement for students involved in ISS, especially for those repeatedly assigned to ISS. The comments included:

"I would staff it 100% of the time. I understand the reason it is not staffed, but it would be much more effective if it were staffed" (Interview, 1/24, #8).

"I would have a monitor so students wouldn't lay their heads down and go to sleep" (Interview, 1/18, #2).

"By having someone in there, it would take the pressure off the principal and the assistant principal" (Interview, 1/24, #8).

"If you had someone who was running it all the time the students wouldn't be sleeping and doing things in there they shouldn't" (Interview, 1/25, #11).

Most teachers stated that the ISS program helped them. These teachers felt that the existence of the ISS program provided another strategy for control of discipline in the classroom. There was not clear consensus among the teachers on the reduction of tardies by ISS; one day in ISS is the usual punishment received by students having four consecutive tardies in one class. Five teachers stated that ISS had reduced tardies while three teachers said ISS had no effect on tardies. Overall teachers felt that the ISS program had changed student behavior at school and in classrooms. Comments included:

"It allows me to remove the extreme problems" (Interview, 1/24, #9).

"If you send them once, they don't want to go back. If they misbehave you can threaten them and they will generally settle down. I can threaten any student with ISS and it will have some influence on them" (Interview, 1/23, #5).

"Students that have been in ISS have quit fighting and grades have come up" (Interview, 1/25, #12).

"ISS teaches and reinforces respect and responsibility. These are two major problems with students today, a lack of respect and responsibility" (Interview, 1/25, #12).

Student Perceptions

Student questionnaires were given to all students attending English classes over a period of two days. A total of 233 students responded to the open-ended questions concerning the ISS program. Of the students responding to the questionnaires, 47 students had been in ISS one time, 40 students had been in ISS more than once, and 146 students had never been in ISS. Student responses were analyzed for recurring themes, differences in perceptions related to student involvement in the ISS program, and for perceived changes in student behavior resulting from the program.

Of those students that had been in ISS, more than half stated that the program had changed their behavior. Students that had been in the ISS program only once showed a higher percentage of stated behavior change (32 students of 47, 68%) than students that had been in the program more than once (22 students of 40, 55%). Students punished with ISS more than once stated more frequently that ISS had little impact on their behavior. This is a common finding which suggests that ISS has less impact on the chronically disruptive student (Siskind & Others, 1993). Stated reasons for changes in behavior as a result of ISS included:

"Because I don't like to stay in there all those hours doing work" (#O22).

"Because I used to show out. I mean get in a lot of trouble. But now my attitude has changed because it's not a fun place to be" (#O12).

"Because I try very, very hard so I won't have to go there anymore" (#O7).

"I have not been as disruptive or been late as much as I used to be" (#O6).

"Because ever since the last time I was in ISS I have been making better grades and I don't talk in class as much as I used to" (#O5).

"I will not be tardy for class anymore" (#O30).

Negative comments about the ISS program changing behavior included:

"I've gotten worse because I know that the worst that can happen is getting sent there" (#O26).

"D.R. is like a sleeping day at school" (#O3).

"Because I didn't go for conduct in the first place. I went for being late too many times" (#Y36).

Students never in ISS strongly believed that it had not changed their behavior because their behavior did not need changing. There were a few comments from these students concerning changes in behavior:

"I try not to be tardy all the time and get in trouble because I would be bored in in-school suspension all day long" (#B48).

"It has made me more careful and makes me make better decisions" (#B78).

Almost all students surveyed expressed negative feelings regarding being in the ISS room. Also, some students expressed feelings of embarrassment towards ISS. Students' comments included:

"I don't ever want to have to go to in-school-suspension" (#B6).

"It is very bored, and too quiet just be doing work whole time the camera be watching you" (#Y38).

"I don't want to go, but I think it's a good thing" (#B48).

". . . people are always saying they don't want to go to DR because it's dead" (#Y39).

Students' responses to the purpose of ISS reflected statements made by administrators and teachers. All participants (students, teachers, and administrators) felt that ISS provided an alternative to out-of-school suspension that provided real punishment for misbehavior. Students' comments included:

"I think the purpose of in-school suspension is to try and give the students a chance to get their act together before they send them home" (#B11).

"To punish people without giving them a vacation" (#B10).

"Students get punished without being sent home" (#O34).

"In-school-suspension is a way teachers know you are not on the streets and you are still having to come to school because some people get in trouble just so they can go home but with in-school-suspension you still have to come to school" (#O6).

Many students never in ISS felt that the program removed the disruptive students so that other students could continue to learn. Some students never in ISS also felt it was not as severe as a typical three day out-of-school suspension. These students felt that ISS was not an effective punishment because ISS students tended to repeat their disruptive behavior. A few students also perceived it as less severe because parents were not informed when students were sent to ISS. Students' comments included:

"To keep unruly children from distracting others" (B#35).

"The purpose of in-school suspension is to get the students out of the teacher's room so they will not have to put up with those students" (#B3).

"For the ones who don't follow the rules. The ones who misbehave in the classroom" (#B12).

"Because some students have been in there a lot and the teachers have not had to put up with them" (#B3).

Many students surveyed, both those who had and had not been in ISS, expressed that frequently students were going into ISS to get out of regular classes and to sleep. Sleeping while in ISS was viewed as a primary problem by all participants in the study: administrators, teachers, and students. Student comments included:

". . . you can sleep all day in there, no one cares. If your (sic) not acting up they leave you alone" (#O11).

". . . the only thing students do is sleep, and they feel its just a place to blow time and take a quick nap" (#O36).

"When someone is in there we don't do anything but sleep" (#Y24).

". . . some people get in there to get out of classes" (#Y12).

Observations of the ISS Room

The ISS room in this study is unique in that it does not use a staff person in the room to direct students' activities. Students are observed continually by the secretary, assistant principal, or principal using a camera and monitor. Students' activities can be seen and heard in great detail over the monitor. Students that are not working or performing as instructed are redirected by the principal or the assistant principal.

The observation camera and monitor were used to determine the type of student behavior taking place in the ISS room. Observations of the ISS room were made over a four day period, at different times throughout the school day. During this period 63 student behaviors were observed, separated into three categories, and counted to determine the frequency of each category of student activity. Student working behaviors made up 38 of the 63 (60%) of total student behaviors. The remaining 40% non-working student behaviors included: 12 (19%) students doing something other than work or simply sitting doing nothing; 13 (21%) students resting their heads on their desks.

Five students exhibiting this head down behavior were questioned by the researcher as to the reason for the behavior. Four students explained they were out of work and had nothing to do. One student was simply not doing the work on his desk. No supplemental books or activities were available in the ISS room for students to use once all class work was completed.

Conclusion

Teachers and students perceived that the new ISS program positively affected school discipline. Both groups viewed the program as a positive alternative to out-of-school suspension.

During this initial year of ISS implementation, teachers found the program helpful with classroom discipline. All teachers and many students viewed ISS as an effective way to remove disruptive students from the classroom. Over half the students that had attended ISS stated the program had changed their behavior. These students viewed the ISS room as unpleasant and stated that their behavior had changed as a result of not wanting to return. Almost all students indicated that they did not want to be in the ISS room. Although the findings of the study were generally positive, the researchers recommend further in-depth study incorporating extended observation periods to better determine the long-term effectiveness of the program on classroom behavior.

The largest problem perceived by both teachers and students as negatively impacting program effectiveness was the frequent occurrence of students sleeping while in ISS. Another often cited problem was the high number of students who repeatedly returned to ISS.

This evaluation of the initial year of the ISS implementation resulted in varied findings. The program failed to reduce the number of lost instructional days and the number of out-of-school suspensions. No records were maintained regarding the number nor type of ISS incidences, therefore, no conclusions could be made concerning recidivism. Students in ISS were offered no counseling to address their behavior, and no behavior monitoring was conducted after ISS. In contrast to these findings, administrators, teachers, and students did perceive that ISS had positively impacted student discipline and reduced the number of incidences of student misbehavior.

The purpose of the camera-monitored ISS program was to provide an alternative to out-of-school suspension without incurring the expense of a full-time monitor. The authors conclude that this type of program can be successful for some misbehavior. Both camera-monitored as well as staff-monitored ISS can address behaviors which include smoking, tardies, profanity, and skipping class. Also public displays of affection, repeated lack of homework, dress code violations, and misbehavior in the cafeteria or on the school bus can be handled effectively with ISS measures.

When developing an ISS program, the authors recommend that all offenses resulting in ISS as well as the program's policies and procedures be clearly stated in student, parent, and teacher handbooks. Since ISS should not be considered a place to retain disobedient students, it should be communicated that the objective of the program is to help young people to learn to exercise socially acceptable behavior. Therefore, the number of times a student can be assigned to ISS should be limited to 2 or 3 assignments per grading period with parent notification for each assignment. Afterward, disciplinary measures would escalate to more severe actions such as out of school suspension.

Some schools require a teacher to provide a student numerous warnings prior to ISS assignment. The authors discourage this practice in that it provides an unnecessary hardship for teachers and allows misbehaving students to remain in the class. A reasonable number of warnings such as two or three is advised.

For a successful program, total isolation is essential. Therefore, ISS should begin as soon as the student enters school grounds and end when the student is dismissed. If possible, the ISS room should be isolated from the main building and minimal in visual and auditory distractions. This can be accomplished by obtaining a temporary building and replacing its desks with cubicles. Students should be required to complete class assignments, homework, and additional supplementary tasks available in the ISS room if time permits. Students must be constantly engaged in class and supplementary assignments. Off-task behavior, peer interaction, and sleeping should be strictly prohibited. Lunches and restroom breaks may occur when other students are in class.

After students leave ISS, their behavior should be monitored closely to reinforce positive behavioral changes. Furthermore, a strong counseling component is essential. Counseling should take place during ISS and continue during the post-monitoring process. Although a full-time monitor may not be feasible, counseling should be provided to each student during and after ISS to ensure program success. Accurate records must be maintained to evaluate the effect of the program on recidivism and the reduction of student misbehavior. Most importantly, teachers should receive training on the program and its procedures and be consistently encouraged to enforce the program by not allowing students who commit ISS offenses to remain in the classroom.

The authors conclude with teachers that ISS rooms are best when staffed by full-time knowledgeable and skillful personnel. Unfortunately, rural schools need disciplinary alternatives and often have limited resources. Although not preferred, a camera-monitored ISS program incorporating accurate record keeping, proper monitoring, appropriate on-task assignments, and intensive counseling during and after ISS can be used effectively to address many disciplinary problems in small rural schools.

References

Billings, W. H. & Enger, J. M. (1995, November). *Perceptions of Missouri high school principals regarding effectiveness of in-school suspension as a disciplinary procedure*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Biloxi, MS. (ERIC Document Reproduction Service No. ED 392 169)

Corbett, A. H. (1981). Is your ISS program meeting its goals? Take a closer look. *NASSP Bulletin, 65* (448), 59-63.

Corbett, W. T. (1990). The time-out room: A key component of a total school discipline program. *The Clearing House, 63*, 280 - 281.

Disciullo, M. (1984). In-school suspension: An alternative to unsupervised out-of-school suspension. *The Clearing House, 57*, 328-330.

Fischel, F. J. (1986). In-school suspension programs - questions to consider. *NASSP Bulletin, 70* (493), 100-102.

Jones, R. R., Jr. (1983). Sorry, partner, but your suspension is here at school. *Principal, 62* (5), 42-43.

Mendez, R., & Sanders, S. G. (1981). An examination of in-school suspension: Panacea or Pandora's box? *NASSP Bulletin, 65* (441), 65-69.

Montgomery County Public Schools. (1981). *A preliminary evaluation of the pilot in-school suspension program, 1980-81*. Rockville, MD: Department of Educational Accountability. (ERIC Document Reproduction Service No. ED 228 359)

Silvey, D. F. (1995). *The effect of inschool suspension on the academic progress of high school science and English students*. (ERIC Document Reproduction Service No. ED 389 069)

Siskind, T. G. (1993, February). *An evaluation of in-school suspension programs*. Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater Beach, FL. (ERIC Document Reproduction Service No. ED 360 718)

Sullivan, J. S. (1989). Elements of a successful in-school suspension program. *NASSP Bulletin, 73* (516), 32-38.

Weiss, K. (1983). In-school suspension - time to work, not socialize. *NASSP Bulletin, 67* (464), 132-133.

Whitfield, D. & Bulach, C. (1996, March). *A Study of the effectiveness of an in-school suspension*. Paper presented at the National Dropout Prevention Network Conference, Tampa, FL. (ERIC Document Reproduction Service No. ED 396 372)

Yin, R. K. (1994). *Case study research design and methods* (2nd ed.). Thousand Oaks, CA: Sage.

Copyright 1997 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1997, Vol. 4, No. 2, 65-70

Score Comparisons of ACCUPLACER (Computer-Adaptive) and COMPANION (Paper) Reading Tests: Empirical Validation and School Policy

Jason C. Cole
California School of Professional Psychology, San Diego, California

Anthony D. Lutkus
Educational Testing Service, Princeton, NJ

This study highlights the importance of empirically validating hypotheses in an academic setting. Herein, the authors compared the Reading Comprehension test scores of entering college students who took the computer-adaptive ACCUPLACER and its parallel, paper-and-pencil COMPANION form. A client college gave the computerized placement test to 399 students and the paper-and-pencil version to 481 students. Though results showed a significant difference between the two groups, non-random, biased assignment techniques had been employed. Although the institution initially doubted the equivalency of the two forms of the Reading Comprehension test, once the age of the two groups was held constant no difference between the two groups of subjects remained.

ACCUPLACER (College Board, 1995a) is a software system designed to provide basic skills placement, advisement and guidance information for students entering institutions of higher education. Part of ACCUPLACER utilizes a computerized adaptive test format (CAT; see Wainer, 1983; Ward, 1986), called the Computerized Placement Tests (CPTs). The CPTs are delivered through multiple-choice items. Recently, a paper-and-pencil version of the CPTs, called COMPANION, was developed by selecting items directly from the CPTs' item pool. Educational Testing Service (ETS) used an automated item selection computer program that preserved ACCUPLACER's test specifications. In the Reading Comprehension test, ACCUPLACER's computer-adaptive format delivers 20 questions (items) while COMPANION contains 35 items. Item Response Theory (IRT) parameters are used to equate the two forms. It was assumed that no modality difference existed between the two test delivery formats (see Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991) and the same scale is used for both test forms.

The assumption of equivalency between the forms was recently questioned by one of the users of the ACCUPLACER system, specifically for the Reading Comprehension test. The college presented data from 399 students tested on the CPTs and 481 students tested on the COMPANION.[1] The mean CPTs Reading Comprehension scaled score was 77.51 (*SD* = 20.90). The mean COMPANION Reading Comprehension scaled score was 72.62 (*SD* = 19.02). The college initially believed that the CPTs were easier than the paper-and-pencil format and considered adjusting their placement cut-off scores. Before proceeding with a class placement policy change, the school contacted the authors of this article at ETS for consultation on the apparent lack of form equivalency.

The first aspect of inquiry by the authors regarded the methods used to collect students. Immediately a problem was detected: CPTs students included anyone who approached the college admissions office for testing, while COMPANION scores came only from a mandatory testing at a high school. The authors hypothesized that the data may have shown group differences due to the non-random assignment employed by the college. However, there was still no empirical support for why the differences in scores existed.

Thus, the hypothesis of why the test results differed laid in the methodology of administration of the tests. Subjects were not randomly assigned to the treatment groups – CPTs or COMPANION testing systems. Researchers have long known that there are problems associated with non-random assignments of subjects to treatment groups, yet providing clear-cut evidence of

---

Jason C. Cole, Department of Clinical Psychology; Anthony D. Lutkus, College Entry-Level Assessment Program. Special thanks to the staff at the College Board and Educational Testing Service, especially Michelle Rosenthal, Lisa Dickens, and Howard Everson. The authors would also like to thank Laura Longo at Brookdale Community College for her assistance with the data. Also, we would like to thank Kathleen A. Haley, from Boston College, for her insights and editorial accuracy. Tom R. Smith was also very helpful in the editing – thank you. Last, special sincere thanks, Alan S. & Nadeen L. Kaufman, and Nusheen Cole. Correspondence concerning this article should be addressed to Jason C. Cole, PO Box 2664, Laguna Hills, California, 92654-2664. Electronic mail may be sent via Internet to JasonCCole@home.net.

these problems is somewhat elusive (Keppel, 1991, p. 97). Publishers of psychological and educational test batteries often face perils of similar nature when clients present data that do not support the test's claims. While a review of the assignment-to-groups method used by clients is often a starting point for test publishers, a critique bereft of an empirical answer as to *why* the groups differ in performance often leaves clients feeling their quandary is unanswered.

Method

Subjects

Students for the CPTs consisted of those coming to a specific New Jersey community college for information or registration and desiring to take a placement test. The mean age for this group was 24.4 years, *SD* = 8.33 years. Students tested on the COMPANION were given the test at a high school near the college in a compulsory manner. Demographic data were not available for these students, however, they were all high school seniors. Thus, it was assumed that most of these students were under 18½ years old.

Apparatus

All students were given tests from the ACCUPLACER placement system. The CPTs group was given the Computerized Placement Tests from ACCUPLACER Version 4.5 (College Board, 1995a), while the COMPANION group were given paper-and-pencil COMPANION from the ACCUPLACER system (COMPANION tests do not have different versions). Both the CPTs and COMPANION version contain four core subtests which were given to all subjects. These subtests include two math tests (Arithmetic and Elementary Algebra) and two English Tests (Reading Comprehension and Sentence Skills). All subtest scores were converted to scaled scores, with a possible range from 20 to 120. These scores are estimated "total right" scores, calculated through Item Response Theory estimation. Students given the CPTs used a computerized test administration, which included a tutorial on how to complete the test. Students given the COMPANION were high school students completing an administration of the paper-and-pencil test. A detailed discussion of the psychometrics of the ACCUPLACER system is beyond the scope of this article, but information can be obtained through Educational Testing Service. A detailed review of ACCUPLACER has also been written by Cole (in press).

Currently, there are no studies that have assessed the relationship between the CPTs and COMPANION versions of ACCUPLACER. However, the COMPANION items were drawn from the CPTs' item pool and have the same range of difficulty, content level, sensitivity, gender relevance, and timeliness as the CPTs (College Board, 1995b). Scores from COMPANION subtests are therefore considered to be equivalent to their respective CPTs counterpart.

Procedure

An ANOVA was calculated to confirm the difference in test modality found by the college requesting the analyses. This ANOVA used the test modality (COMPANION and CPTs) as the independent variable and Reading Comprehension scaled score as the dependent variable. Upon confirmation of the modality difference, two separate sets of exploratory analyses were conducted.

The first set of analyses explored the possibility of an overall verbal ability difference between the two groups. An ANOVA was conducted to determine the difference between the two sets of subjects for the Sentence Skills scaled scores. Again, the test modality was used as the independent variable, whereas the Sentence Skills scaled score was the dependent variable. A preliminary correlation between the two subtests (for the sample used in this study) indicated a significant correlation between these subtests: $r = .68$, $p < .001$. This represents a large effect (Cohen, 1992).

A second set of analyses explored the difference between CPTs and COMPANION subjects when age was held constant. Date of birth and date of test were available for the CPTs students, but not for the COMPANION students. Since the COMPANION users were all high school seniors, their age could be assumed to be below 18 and a half years old, with very few exceptions. CPTs subjects who were older than 18.5 were eliminated from the analysis. The remaining 160 CPTs subjects had a mean age of 17.7 years (*SD* = 118 days). Means, *SD*s, and *n*s for the two groups' scores on the Reading Comprehension subtest are located in Table 1. An ANOVA was used to assess the difference between the COMPANION subjects and the CPTs subjects, matching for age. Again, the independent variable was test modality, and the dependent variable was Reading Comprehension scaled score.

Table 1
Descriptive Statistics for the Groupings of CPTs and COMPANION Students

Subject Group and Subgroup	*Mean*	*SD*	*n*
COMPANION Main Group	72.62	19.02	481
CPTs Main Group (without age covaried)	77.51	20.90	399
CPTs Matched Group (Ages 17-18.50)	73.24	20.35	160

A total of two analyses were conducted. Therefore, in order to reduce the probability of a Type I error, a Bonferroni correction was used to adjust the p levels throughout all analyses. Two analyses divided by the standard alpha level of .05 adjusted the alpha level for the analyses contained herein to .025.

It should be noted that technically normality was violated in two analyses in this study. Yet, at least two Monte Carlo studies (Clinch & Keselman, 1982; Tan, 1982) that assessed the violation of normality with the F test have determined that the violations such as those exhibited in this article produce negligible interpretative difficulty of the F statistic: F is robust to the violation of normality. Also, as the violation of normality always *increases* the probability of a Type I error (Keppel, 1991, p. 97), it would appear that a radical violation of normality still would not occlude interpretation of either violated analysis in this article. Keppel suggests lowering the lower alpha level when egregious violation of normality occurs. The first analysis with a violation of normality had a $p < .001$, and the second analysis had a $p > .025$. Therefore, only the first analysis might have been affected by lowering the alpha to control for any Type I error problems. Yet, with a p value of less than .001, even an extremely conservative alpha adjustment would not affect the rejection of the null hypothesis.

Results

The normality of a distribution of scores is assumed with an analysis of variance (Keppel, 1991, p. 97). The means and *SD*s for the CPTs and COMPANION students are presented in Table 1. The distribution of data for the CPTs (skewness z = -4.51, kurtosis z = -1.33) and COMPANION (skewness z = -1.64, kurtosis z = -2.41) scores presented one difficulty: the skewness for the CPTs scores was large. Typically, a skewness or kurtosis z greater than 3 is considered suspect. However, as previously noted, there are at least two Monte Carlo studies (Clinch & Keselman, 1982; Tan, 1982) that demonstrated that skewness discrepancies of this level present minimal interpretive difficulty. In fact, skewed distributions are very common in placement tests. Homogeneity of variance was also assessed. The F_{max} test for homogeneity of variance was 1.21, and showed a nonsignificant difference between the variances of the groups.

An ANOVA on the two groups was performed, emulating the analysis run by the college that collected these data. The independent variable in the ANOVA was the mode of delivery for the test (either computerized or paper-and-pencil), while the dependent variable was the Reading Comprehension scaled score. The groups were significantly different: $F(1, 878) = 13.14, p < .001$. The magnitude of effect (ω^2) was .014 – a small effect (Cohen, 1977).

An analysis of the difference in Sentence Skills test scores between the CPTs and COMPANION students was conducted to assess a difference in verbal ability (writing/editing skills) between the two groups. Normality and homogeneity of variance were assessed for the Sentence Skills scores. Normality for the CPTs (skewness z = -3.85; kurtosis z = -2.63) had a suspect skewness z. The normality for the COMPANION scores (skewness z = -6.1; kurtosis z = .2) had a suspect skewness z as well. Again, these are not damagingly large skewness scores. The assessment of homogeneity of variance produced a nonsignificant F_{max} of 1.15.

An ANOVA was computed with the independent variable as the test modality and the dependent variable as the Sentence Skills scaled scores. A nonsignificant result between the groups resulted: $F(1, 878) = 2.34, p > .025$. Power was adequate (Power > .80) to see down to a small effect, given the number of subjects per group (see Cohen, 1992). In summary, the results showed no overall difference in verbal ability as measured by the Sentence Skills subtest between the CPTs and COMPANION students.

Next, the effect of test modality was assessed by holding the age range constant for the CPTs users – CPTs subjects could be no older than 18.5 years of age. This limited group of CPTs subjects were then compared to the COMPANION subjects for differences in Reading Comprehension scaled scores. Skewness and kurtosis statistics for the two groups are presented in Table 2. Normality was viable for the measures. Homogeneity of variance was also assessed with the largest and smallest variances amongst the groups: the CPTs group (414.18) and the COMPANION group (361.81). F_{max} = 1.15 was nonsignificant, and thus homogeneity of variance existed for these groups.

Table 2
Normality Statistics for the CPTs Subgroups

Statistic	CPTs Age <18.5	COMPANION
Skewness z	-1.99	-1.65
Kurtosis z	-1.51	-2.38

An ANOVA using the age-controlled subjects between the CPTs and COMPANION groups was not significant $F(1, 639) = 0.12, p > .025$. This ANOVA used test modality (CPTs or COMPANION) as the independent variable and Reading Comprehension scaled score as the dependent variable. Power was sufficient to see down to a small effect (see Cohen, 1992).

The above ANOVA result was indicative of a relationship between age and Reading Comprehension scaled score, at least for the CPTs group. To test the assertion, a correlation was run for the CPTs subjects (all 399), correlating ages and scaled scores. There was a significant correlation: $r = .19$, $p < .001$[1]. Cohen (1992) notes that correlations of this size are between a small and medium effect size.

Discussion

The impact of this study assisted in providing the college with important information in order to make well informed policy decisions. Moreover, the global ramifications of this study assist in elucidating the need for empirical support for administrative decisions in learning institutions. The initial and most direct conclusion was that a difference between the computerized and paper-and-pencil test modalities existed. Yet, had the college proceeded to make policy changes regarding class placement on a modality difference assumption, the result might have led to ill-prepared students taking classes they couldn't handle (or adequately prepared students taking classes too easy for them). Overall, the long-term results could have been massive.

An intriguing difference was found during the data exploration in this study; unlike Reading Comprehension, Sentence Skills was not impacted by age. Reading Comprehension and Sentence Skills are both crystallized (g_c) tasks, as described by Horn and Cattell (1966; 1967). Essentially, g_c increases at a linear rate throughout childhood and adolescence. In adulthood, this ability increase attains its apogee where it then remains as a steady plateau (sometimes shown with a slight decline beginning around age 45-50) throughout the rest of one's life. Many previous studies on the relationship between age and crystallized intelligence (g_c) have demonstrated such a relationship (Horn, 1978; Horn & Cattell, 1966; Horn & Donaldson, 1976; Horn & Donaldson, 1980; Horn, Donaldson, & Engstrom, 1981; Kaufman & Kaufman, 1993; Matarazzo, 1972; McGrew, Werder, & Woodcock, 1991; Woodcock, 1978). Therefore, the *differential* impact age exhibited on the Reading Comprehension and Sentence Skills subtests of ACCUPLACER was a bit intriguing.

Horn (1985) later expanded the fluid-crystallized theory to contain many subdivisions of intelligence. Woodcock and Johnson have used Horn's modified fluid-crystallized theory in their revised psychoeducational assessment battery (WJ-R; Woodcock & Johnson, 1989b), noting that their test is "... an operational representation for a particular theory of intellectual processing – the Horn-Cattell g_f-g_c model" (Woodcock & Mather, 1989, p. 13). The authors of the technical manual for the WJ-R provided growth curves for specific achievement tasks, plotting age against W score[2] for different achievement categories (McGrew et al., 1991). A close look at the aforementioned growth curves provided evidence for two pertinent differences amongst the various achievement tests' growth curves: amount of growth during adolescence, and age of maximum performance (curve apogee) varied from test to test. The current authors believed that the differences among the aforementioned growth curves were likely generalizable to the differences between the relationship of age and the two English CPTs subtests.

Woodcock (personal communication, August 10, 1997) noted that the differing impact of age on Reading Comprehension and Sentence Skills scores did not defy the g_f-g_c model. In fact, given that different crystallized subtests may have different age growth curves, Woodcock believed the findings from this study to be reasonable. He further noted the Sentence Skills, a test of grammatical ability, showed an intuitive relationship with age; individuals older than school-age needn't harness these skills much further. Yet, he felt that Reading Comprehension, a measure of one's ability to understand information read, was a vital skill used frequently after one's compulsory education.

Woodcock (1997) also provided validation of the differences exhibited in this study for Reading Comprehension and Sentence Skills. The validation data were obtained from the normative data set of the WJ-R Achievement test (WJ-R/A; Woodcock & Johnson, 1989a). The Proofing subtest from the WJ-R/A is similar to the Sentence Skills subtest from ACCUPLACER. The correlation between the Proofing W score and age was $r = .79$ ($p < .001$, based on an n of 2, 212) for individuals 6 to 16 years-old. Yet, for subjects at least 17 years-old, this correlation dropped to $r = .04$ ($p > .05$, based on an n of 934). Reading Vocabulary on the WJ-R/A, a test similar to Reading Comprehension from ACCUPLACER, was more highly related to age for adults than the Proofing subtest. The correlation between age and Reading Vocabulary W score for children ages 6 to 16 was $r = .78$ ($p < .001$, based on an n of 2, 212): similar to the correlation between Proofing and age for the same age cohort. However, for individuals at least 17 years-old, the correlation between age and Reading Vocabulary was still .18 ($p < .001$, based on an n of 934). The correlation between Reading Vocabulary and age was very similar to the correlation obtained in this study between Reading Comprehension and age ($r = .19$).

This data exploration did not prove the equivalency of the COMPANION and the CPTs reading tests: it did show CPTs scores vary as a function of one's age. Further, for the sample used in this study, the differences between the COMPANION and CPTs scores were negligible, once age was considered. Although the results tend to support equivalency, this was not an equivalency study. Future research should attempt to empirically validate the equivalency of these two test modalities for the ACCUPLACER system.

Some limitations of this study are related to the brevity of information on the subjects. More information would have helped to assess if any other demographic variables were significantly related to scaled scores. It should be noted, ETS has conducted detailed Differential Item Functioning analyses (for further information, see Holland & Thayer, 1988) on the ACCUPLACER system, and any biased items have been eliminated (College Board, 1995b). These analyses, however, were limited to assessing racial and gender differences. The assessment of other demographic information may have shown such differences (for example, socioeconomic status), or overall biasing effects not noted at the item level for race or gender. Another limitation of this study was the lack of specific ages for the COMPANION subjects. Whereas a good argument can be made that nearly all of these subjects would be under 18.5, accurate information is always preferred. Last, the concept that cohort groupings exist for varying means on the CPTs brings forth another question: does the predictive validity of the CPTs vary as a function of age cohort? While predictive validity estimates are provided in the ACCUPLACER manual (College Board, 1995a), there have not been any studies regarding the influence age has on the predictive validity estimates. Future research should explore this area. Although replication of this study's findings may be useful, the authors feel that future research would be most beneficial assessing the impact of age on ACCUPLACER scores, and assessing the comparability of the CPTs and COMPANION.

Summarily, the authors of this study found empirical support for the impact of non-random assignment by the college collecting the comparative data for the CPTs and COMPANION tests. No difference, in fact, was uncovered by the authors between the tests when age was held constant. The difference noted by the college was due to the unequal age distributions of the two groups combined with the medium effect age has on the Reading Comprehension subtest of ACCUPLACER throughout adulthood. However, the other English subtest of ACCUPLACER, Sentence Skills, was not significantly impacted by age for older adults. This discrepancy between two crystallized subtests was explained using Horn's modified g_f-g_c theory of intellectual processing (1985).

References

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics, 7*, 207-214.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* (Rev. ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cole, J. C. (in press). ACCUPLACER: A review of a college entry-level placement system. *Journal of Psychoeducational Assessment.*

College Board. (1995a). *ACCUPLACER: User's Notebook Version 4.5.* New York: College Entrance Examinations Board.

College Board. (1995b). *ACCUPLACER: User's Notebook Version 5.0.* New York: College Entrance Examinations Board.

Holland, P. W., & Thayer, D. T. (1988). Differential performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-146). Hillsdale, NJ: Erlbaum.

Horn, J. L. (1978). Human ability systems. In P. B. Baltes (Ed.), *Life-span development and behavior* (Vol. 1, pp. 211-256). New York: Academic Press.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 267-300). New York: Wiley.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253-270.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129.

Horn, J. L., & Donaldson, G. (1976). On the myth of intellectual decline in adulthood. *American Psychologist, 31*, 701-719.

Horn, J. L., & Donaldson, G. (1980). Cognitive development, II. Adulthood development of human abilities. In O. G. Brim & J. Kagan (Eds.), *Constancy and change in human development: A volume of review essays* (pp. 445-529). Cambridge, MA: Harvard University Press.

Horn, J. L., Donaldson, G., & Engstrom, R. (1981). Apprehension, memory, and fluid intelligence decline in adulthood. *Research on Aging, 3*, 33-84.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test.* Circle Pines, MN: American Guidance Service.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Matarazzo. (1972). *Wechsler's measurement and appraisal of adult intelligence*. (5th ed.). New York: Oxford University Press.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP examinations* (College Board Report 91-5). Princeton, NJ: Educational Testing Service.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *WJ-R technical manual: A reference on theory and current research*. Itasca, IL: Riverside Publishing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communication in Statistics -- Theory and Methods, 11*(2), 485-511.

Wainer, H. (1983). On item response theory and computerized adaptive tests: The coming technological revolution in testing. *The Journal of College Admissions, 27*, 9-16.

Ward, W. C. (1986). *Using microcomputers to administer tests*. Princeton, NJ: Educational Testing Service.

Woodcock, R. W. (1978). *Development and standardization of the Woodcock-Johnson Psycho-Educational Battery*. Allen, TX: DLM/Teaching Resources.

Woodcock, R. W. (1982, March). *Interpretation of the Rasch ability and difficulty scales for educational purposes*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Woodcock, R. W. (1997). [Achievement correlations for children and adults]. Unpublished raw data.

Woodcock, R. W., & Dahl, M. N. (1971). *A common scale for the measurement person ability and test item difficulty* (AGS Paper No. 10). Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. (1989a). *Woodcock Johnson Tests of Achievement: Standard and Supplemental Batteries*. Chicago, IL: Riverside Publishing.

Woodcock, R. W., & Johnson, M. B. (1989b). *Woodcock-Johnson Psycho-Educational Battery -- Revised*. Chicago, IL: Riverside Publishing.

Woodcock, R. W., & Mather, N. (1989). WJ-R Tests if Cognitive Ability -- Standard and Supplemental Batteries: Examiner's manual. In R. W. Woodcock & M. B. Johnson (Eds.), *Woodcock-Johnson Psycho-Educational Battery - Revised*. Allen, TX: DLM Teaching Resources.

Wright, B. D. (1979). *Best test design*. Chicago, IL: MESA Press.

Footnotes

[1] Subjects with any missing data were eliminated from all analyses. The percentage of subjects missing data did not exceed 2% of the total number of subjects.

[2] A W score is a Rasch based ability score (for further information, see Rasch, 1960; Woodcock, 1982; Woodcock & Dahl, 1971; Wright, 1979).

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◇◇

Please enter a subscription to RESEARCH IN THE SCHOOLS for:

Name: ____

Institution: ____

Address: ____

____

			COST
Individual Subscription ($25 per year)	Number of years	_____	________
Institutional Subscription ($30 per year)	Number of years	_____	________
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	_____	________
Back issues for Volumes 1, 2, 3, and 4 ($30 per Volume)	Number of Volumes	_____	________
TOTAL COST:			________

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
RESEARCH IN THE SCHOOLS
University of Alabama at Birmingham
School of Education, 233 Educ. Bldg.
901 13th Street, South
Birmingham, AL 35294-1250

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂✂

MSERA Membership/Renewal Form

(Please print or type)

Name ____

Organization ____

Address ____

____

Telephone Work: ____________________

Home: ____________________

Fax: ____________________

e-mail: ______________________________

Amount Enclosed:	MSERA 1998 Membership ($25 professional, $15 student)	$______
	MSER Foundation Contribution	$______
	TOTAL	$______

Make check out to MSERA and mail to:

Dr. Clifford Hofwolt
MSERA Secretary-Treasurer
Vanderbilt University
Box 330, Peabody College
Nashville, TN 37203

RESEARCH IN THE SCHOOLS
Mid-South Educational Research Association
and the University of Alabama at Birmingham
901 South 13th Street, Room 233
Birmingham, AL 35294-1250

BULK RATE
U.S. POSTAGE
PAID
PERMIT NO. 1256
BIRMINGHAM, AL

MSERA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and the University of Alabama at Birmingham.

Volume 5, Number 1 **Spring 1998**

Dropping Out of Secondary School: A Descriptive Discriminant Analysis of Early Dropouts, Late Dropouts, Alternative Completers, and Stayins 1
Todd C. Campbell and Michael Duffy

Applications, Trials, and Successes of CU-SeeMe in K-12 Classrooms in the Southeast 11
Anna C. McFadden, Margaret L. Rice, Vivian H. Wright, Roger Saphore, and Jenka Keizer

Secondary School Size and Achievement in Georgia Public Schools 18
Ellice P. Martin and John R. Slate

Fourth and Fifth Grade Students' Attitudes Toward Science: Science Motivation and Science Importance as a Function of Grade Level, Gender, and Race 27
John R. Slate and Craig H. Jones

Quantitative Graphical Display Use in a Southern U.S. School System 33
John V. Dempsey, Samuel H. Fisher, III, and Judith B. Hale

Statistics Anxiety: A Function of Learning Style? 43
Anthony J. Onwuegbuzie

Topic Coverage in Statistics Courses: A National Delphi Study 53
Kathleen Cage Mittag and Elizabeth M. Eltinge

Modeling Asymmetric Hypotheses with Log-Linear Techniques 61
Frank Lawrence, Gerald Halpin, and Glennelle Halpin

James E. McLean and Alan S. Kaufman, Editors

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250. Please direct questions to jmclean@uab.edu**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1998 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 5, Number 1 **Spring 1998**

EDITORS
James E. McLean, *University of Alabama at Birmingham*
and Alan S. Kaufman, *Yale University School of Medicine*

PRODUCTION EDITOR
Margaret L. Rice, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Gypsy A. Abbott, *University of Alabama at Birmingham*
Charles M. Achilles, *Eastern Michigan University*
Mark Baron, *University of South Dakota*
J. Jackson Barnette, *The University of Iowa*
Larry G. Daniel, *The University of Southern Mississippi*
Donald F. DeMoulin, *University of Tennessee-Martin*
Daniel Fasko, Jr., *Morehead State University*
Tracy Goodson-Espy, *University of North Alabama*
Glennelle Halpin, *Auburn University*
Toshinori ISHIKUMA, *University of Isukuba (Japan)*
JinGyu Kim, *Seoul National University of Education (Korea)*
Jwa K. Kim, *Middle Tennessee State University*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Tech. Community College*
Jerry Mathews, *Auburn University*
Peter C. Melchers, *University of Cologne (Germany)*
Claire Meljac, *Hospital Sainte-Anne (France)*
Soo-Back Moon, *Catholic University of Hyosung (Korea)*
Arnold J. Moore, *Mississippi State University*
David T. Morse, *Mississippi State University*
Jack A. Naglieri, *Ohio State University*
Sadegh Nashat, *The Tavistock Centre (London)*
Anthony J. Onwuegbuzie, *Valdosta State University*
William Watson Purkey, *The University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Georgia Southern University*
John R. Slate, *Valdosta State University*
Scott W. Snyder, *University of Alabama at Birmingham*
Bruce Thompson, *Texas A & M University*

GRADUATE STUDENT EDITORIAL BOARD
Margery Arnold, *Texas A & M University*
Vicki Benson, *The University of Alabama*
Alan Brue, *University of Florida*
Brenda C. Carter, *Mississippi State University*
Jason C. Cole, *California School of Professional Psychology - San Diego*
Robin A. Cook, *Auburn University*
Harrison D. Kane, *University of Florida*
James Kaufman, *Yale University*
Kevin M. Kieffer, *Texas A & M University*
Pamela A. Taylor, *Mississippi State University*
Sherry Vidal, *Texas A & M University*

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 1-10

Dropping Out of Secondary School: A Descriptive Discriminant Analysis of Early Dropouts, Late Dropouts, Alternative Completers, and Stayins

Todd C. Campbell
Marquette University

Michael Duffy
Texas A&M University

Data from the National Educational Longitudinal Study of 1988 (NELS:88) were used to investigate the dynamics of dropping out of secondary school. The consequences of dropping out of high school continue to become more severe both for the individual and for society. Factors affecting high school drop out are likely to present a series of interactive forces ranging from the individual to larger societal levels. These factors can be divided into five major categories: (a) individual, (b) family, (c) peer, (d) community, and (e) school. Descriptive discriminant analyses (DDA) were employed to determine the underlying structure that discriminated between the various groups reflecting educational status in 1992 (i.e., alternative completers, early dropouts, late dropouts, vs stayins or stayins vs dropouts). National, longitudinal, large-sample data were employed to understand better the dynamics of dropping out of secondary school.

The consequences of dropping out of high school have become severe at both a societal and a personal level. Today, the high school dropout is not easily absorbed into the workforce due to ever increasing demands for highly trained workers (Rumberger, 1987). Students who drop out of high school will lack the necessary skills to participate in a high-tech job market and are likely to be destined to marginal employment or outright dependence upon family and/or society (Catterall, 1985). In turn, society will be adversely affected due to the loss of human capital. High school dropouts tend to earn lifetime incomes that are substantially lower than those who graduate from high school, thus reducing the nation's productive capacity (Catterall, 1985; Rumberger, 1983). High school dropouts are likely to receive more welfare benefits, require more health care and unemployment subsidies, and are involved in more criminal activity resulting in increased economic and social burdens upon society (Catterall, 1985).

At the personal level, research indicates that dropping out of high school has negative effects on subsequent cognitive performance and psychological function (e.g., anxiety, depression, self derogation) (Bachman, Green, & Wirtanen, 1971; Kaplan, Damphousse, & Kaplan, 1994). These negative effects can be recursive in nature and perpetuate the disadvantaged position of the high school dropout. However, other studies have shown that dropping out of school may actually benefit some students (Rumberger, 1983). For example, Wehlage and Rutter (1986) found that in comparison to high school graduates, dropouts showed equal or greater improvements in self-esteem and external locus of control.

As Roderick (1993) stated, "Early school leaving is probably the most widely studied educational problem in America. In the 1980s alone, hundreds of books and articles were written on the topic of high school dropouts" (p. 26). However, few national studies have been undertaken and of these only the National Educational Longitudinal Study of 1988 (NELS:88) (Ingels, Abraham, Karr, Spencer, Frankel, & Owings, 1990) has included early dropouts. Also, with rare exception (e.g., Catterall, 1986; Ekstrom, Goertz, Pollack, & Rock, 1986; Pallas, 1985; Rumberger, 1987, 1995), research pertaining to high school dropouts has tended to utilize solely descriptive statistics (primarily percentages) or has focused on bivariate correlations between dropping out and various demographic, individual, family, and school factors (Catterall, 1986; Rumberger, 1987). These methodological practices are prevalent in this area of research despite the pervasive recognition that the act of dropping out of school is a culmination of many varied

---

We gratefully acknowledge the assistance of Professors Bruce Thompson, Ludy T. Benjamin Jr., and Michael Ash for their helpful and insightful feedback in writing this article. Todd C. Campbell received his Ph.D. in Counseling Psychology this past fall from Texas A&M University. He is currently Assistant Professor in the Counseling and Educational Psychology Program at Marquette University. Michael Duffy is Professor of Counseling Psychology at Texas A&M University. Requests for reprints should be sent to Todd C. Campbell, Department of Counseling and Educational Psychology, Schroeder Health Complex, 120, P.O. Box 1881, Milwaukee, WI 53201-1881 or e-mailed to todd.campbell@marquette.edu.

623

processes that have been developing for many years in the lives of those students who drop out, and that these dynamics involve an ecology that is intrinsically multivariate (Bachman et al., 1971; Catterall, 1986; Pallas, 1985; Rumberger, 1987, 1995).

A pervasive problem in the research literature pertaining to dropping out of school is the over-reliance on statistical significance testing. Statistical significance testing is fraught with many problems (Cohen, 1994; Thompson, 1993, 1997a). These problems include the fact that achieving statistical significance is primarily an artifact of sample size. Considering the relatively large sample sizes in many of the studies investigating school dropouts (particularly the national studies), it is not at all surprising that all findings turn out to be statistically significant. This reliance on statistical significance blurs the differentiation between statistically significant results and truly important (i.e., "practical") results. Consequently, it is also important to maximize the fit of a model of reality and an analytic model by not relying too much on analytic tests of statistical significance.

The causes of high school drop out are varied and complex. Single predictors of high school drop out such as SES, race, and urbanicity, do not address the complexity of the phenomenon adequately. In order to investigate the phenomenon effectively, the longitudinal interactions of many variables must be considered. The present study investigated the complex relationships between individual, family, peer, school, and community related factors that lead to the act of dropping out of school.

Factors Associated with Dropping Out

Many different factors have been associated with predicting and mitigating the act of dropping out of high school. Factors affecting drop out are likely to present a series of interactive forces ranging from the individual to larger societal levels. These factors can be divided into five major categories: (a) individual, (b) family, (c) peer, (d) community, and (e) school (Catterall, 1986; Ekstrom et al., 1986; Pallas, 1985, 1987; Rumberger, 1983, 1987, 1995).

Individual characteristics of the dropout

Individual characteristics of dropouts have been the most widely investigated constructs associated with dropping out of high school. Despite the abundance of research focusing on the characteristics of high school dropouts, there is still substantial disagreement as to the efficacy of particular characteristics in predicting and/or mediating dropping out of school. Because many of these individual characteristics are unalterable (e.g., race, gender), in isolation these characteristics lack utility in helping to resolve the dropout issue (Schulz, Toles, & Rice, 1987). Considering the complexity of the process of dropping out of high school, not only must alterable variables be found, but the interactions of these variables must be considered if effective prevention and intervention practices are to be utilized.

Certain individual characteristics have consistently been investigated in relation to dropping out of school (Barro & Kolstad, 1987; Deschamps, 1993; Finn, 1989; Pallas, 1985; Rumberger, 1983, 1987, 1995; Sewell, Palmo, & Manni, 1981). These characteristics are (a) demographics, (b) personality variables, (c) social adjustment and behavior, (d) academic performance, and (e) accelerated role transitions (the early assumption of such adult roles as worker, spouse, or parent).

Family variables

Systemic factors associated with families are integral to understanding the dynamics of dropping out of school. Noteworthy family-related factors associated with dropping out are family composition (Ekstrom et al., 1986; Finn & Owings, 1994; Rumberger, 1983; Zimiles & Lee, 1991), family size (Barro & Kolstad, 1987), parental education level (Baker & Stevenson, 1986; Barro & Kolstad, 1987; Ekstrom et al., 1986; Rumberger, 1983), family mobility (Kaufman, Bradby, & Owings, 1992), little or no learning materials in the home (Rumberger, 1983, 1987, 1995), lack of positive learning experiences in the home, and a non-English speaking home environment (Peng & Lee, 1992). Rumberger (1995) emphasized the need for further investigation of family *processes* such as parental involvement in children's school activities and parenting practices.

Socioeconomic status (SES) is one of the factors most strongly associated with dropping out of high school. A strong negative correlation between SES and dropping out of high school (i.e., as SES increases the likelihood of dropping out decreases) has been found in various studies (Bachman et al., 1971; Barro & Kolstad, 1987; Finn, 1989; Mare, 1980; Rumberger, 1983, 1995).

Though SES has been consistently shown to be negatively correlated with dropping out of school, several studies have presented confounding findings that an increasing number of middle-class students are dropping out of school (Franklin, McNeil, & Wright, 1990; Franklin & Streeter, 1990).

Peers

There is a long history and a large body of literature pertaining to the influences that peers have upon adolescent behavior (Cooley, 1902; Gavin & Furman, 1989; Jessor & Jessor, 1977), but few have examined the influence of peers upon the process of dropping out of school (Ekstrom et al., 1986; Pallas, 1985; Rumberger, 1987, 1995). Many dropouts have friends who are also dropouts, but this is not surprising considering that peer groups in general tend to share similar educational aspirations. Further research regarding the influences of friends and peer groups upon the process of dropping out is needed.

Urbanicity/Geographic region

High school dropout rates tend to be higher in the central cities as compared to suburban and nonmetropolitan areas (Frase, 1989; McMillen, 1992). Geographic regions of the United States also differ in their respective high school dropout rates (Frase, 1989). When considering urbanicity and geographic region as factors in dropping out of high school, the interactions of SES and race/ethnicity must also be considered. For example, the Northeast tends to have lower dropout rates but is the most urban area of the country. It is plausible that the SES of the Northeast region mediates the effects of its urbanicity.

Schools

School factors in relation to the student (e.g., academic achievement, behavioral problems) have received much attention in the research literature. However, little attention has been paid to characteristics of the *schools* per se that are associated with dropping out of high school (Rumberger, 1987, 1995; Wehlage, 1987, 1989). This perspective proposes that certain school characteristics such as school climate, size, location, public vs private, teacher interest in students, and disciplinary practices influence the process of dropping out (Bryk & Thum, 1989; Catterall, 1986; Ingels, Dowd, Baldridge, Stipe, Bartot, Frankel, & Quinn, 1994; Mann, 1987; Ingels, Dowd, Stipe, Baldridge, Bartot, Frankel, & Quinn, 1994; McDill, Natriello, & Pallas, 1987; Wehlage & Rutter, 1986; Weis, Farrar, & Petrie, 1989).

Multivariate Models

There exists a huge body of research pertaining to high school dropouts, but the focus of this research has been primarily aimed at the individual student and analyzed from the univariate or bivariate perspective. Fortunately, there is a progressive movement toward the development of systemic, multivariate, multi-perspective, longitudinal, process-oriented research models that more accurately reflect the "reality" of the complex process of dropping out of school (Thompson, 1994).

Several researchers (Catterall, 1986; Ekstrom et al., 1986; Pallas, 1985; Rumberger, 1995) have proposed and/or investigated multivariate models of dropping out of school. Though singularly these models do not fully represent a systemic, multivariate, multi-perspective, longitudinal, process-oriented research model, a synthesis of these models does reflect the "reality" of the complex process of dropping out of school by including individual, family, peer, community, and school factors (Campbell, 1997). For the present study, variables from the NELS:88 second follow-up study (student and dropout components) that were considered to be representative of (a) individual educational expectations, (b) student/dropout self concept, (c) student/dropout social adjustment and behavior, (d) student/dropout extracurricular and leisure activity, (e) parental expectations for child's education, (f) parental involvement, (g) family transitions, (h) peer educational expectations, (i) peer attitudes, (j) school climate, and (k) perceptions of teachers were selected for analysis.

Method

Participants

The total number of sample members included in this analysis was 9419. There were more females (54%) than males (46%). The sample ethnic composition was 73.0% White (not Hispanic), 10.8% Hispanic, 8.1% Black (not Hispanic), 7.2% Asian Pacific Islander, and 0.9% American Indian. The majority of the sample members were born in 1974 (66.0%), 25.5% were born in 1973, 2.9% were born in 1972 or before, and 1.2% were born in 1975 or later.

Procedures

The National Educational Longitudinal Study of 1988 (NELS:88) (Haggerty, Dugoni, Reed, Cederlund & Taylor, 1996; Ingels et al., 1990; Rock & Pollack, 1995) was designed and sponsored by the National Center for Education Statistics (NCES): U.S. Department of Education. The NELS:88 provided a longitudinal survey of a nationally representative sample of U.S. middle and high school students, including dropouts. Data collection began in 1988 when sample members were in the eighth-grade and participants were subsequently surveyed every two years. Thus the student dropout status data for the study were drawn from the NELS:88 second follow-up study (1992) when most of the sample members were in their senior year of high school (Ingels, Dowd, & Baldridge et al., 1994; Ingels, Dowd, & Stipe et al., 1994).

The same data were collected for all participant groups during the base year (1988). However, in the subsequent "follow-up" data collection phases (the first one two years after the base year, the second four years after the base year), data were collected separately for the "dropout" and "student" groups.

Because the follow-up surveys for the students and dropouts were not exactly the same, only survey items that were administered in both the student and dropout questionnaires were considered in the present study. To create a data set that contained items administered to both student and dropout sample members (allowing for comparative analyses of the groups) a given dropout questionnaire item was merged with the corresponding student questionnaire item to create a new variable that combined dropout and student responses (Campbell, 1997). Any participant with missing data was excluded from the analysis.

For all of the surveys, items that yielded non-interval data were identified and transformed to dichotomous form. This transformation allowed these data to be considered intervally scaled for the statistical analyses.

Discriminant Analysis

Discriminant analysis is a statistical analysis that determines a set of weights (discriminant function coefficients analogous to beta weights in multiple regression) to assign to individual scores so that the ratio of the between groups sums of squares and cross-products of pooled within-group sums of squares will be maximized. This procedure maximizes the discrimination between groups (Fisher, 1936; Huberty, 1994; Huberty & Barton, 1989; Klecka, 1980; McLachlan, 1992). Discriminant analysis can be divided into two distinct methods that are distinguished by the purpose of the analysis. These two methods are: (a) descriptive discriminant analysis (DDA) and (b) predictive discriminant analysis (PDA). The purposes of DDA and PDA are quite different, therefore the statistics used to interpret the results from the two methods differ (Huberty, 1994; Thompson, 1995).

DDA analyses yield statistics measuring degree of differences on the response variables as a function of group membership, and statistics indicating on which response variables the groups most differ. The interpretation of DDA results is approached in a two-stage hierarchical analysis, as are most analyses in the General Linear Model (Thompson, 1997a).

The first question is, "Do I have anything?". To address this question the researcher can consult some combination of information evaluating (a) statistical significance, (b) result effect size, and (c) result replicability. Only if on some basis a judgment is made that the results are noteworthy, does the researcher then ask, "Where does my effect size originate?". Here the researcher consults both standardized weights and structure coefficients (Thompson, 1997b). Response variables not reflecting group differences have near-zero values for both coefficients; otherwise one or both coefficients will be largely non-zero.

Results

Descriptive Discriminant Analysis (1992 Follow-up Data)

Descriptive discriminant analysis was employed in the present study to evaluate the effects of group membership (i.e., alternative completer, early dropout, late dropout, stayin) on response variables (e.g., contemporaneous or subsequent expectations for educational attainment, feelings of self-worth).

The discriminating groups were (a) alternative completers ($n = 269$), (b) early dropouts ($n = 110$), (c) late dropouts ($n = 362$), and (d) stayins ($n = 8678$). Three functions (k-1) were derived from this DDA. The first Wilks' lambda was 0.423701 and was statistically significant ($p < .0001$). The second Wilks' lambda was 0.934356 and was statistically significant ($p < .0001$). The third Wilks' lambda was 0.976924 and was statistically significant ($p < .0001$).

The canonical correlation coefficients for Function I, Function II, and Function III equaled 0.7393, 0.2087, and 0.1519, respectively, making the squared canonical correlations equal to 0.5466, 0.0435, and 0.0231. The squared canonical correlation coefficients indicate that 54.66% of the variance in the response variables was explained by Function I, 4.3% of the variance across the groups was explained by Function II, and 2.3% of the variance was explained by Function III.

The eigenvalues for the three functions were 1.2052, 0.0456, and 0.0236, respectively. Therefore, the first function was 26.43 times better at discriminating between the groups than the second function and 51.07 times better at discriminating between the groups than the third function. Considering the squared canonical correlation coefficients and the eigenvalues for the three functions, only Function I was deemed substantively meaningful in explaining the difference between the groups.

Table 1 presents the standardized function coefficients and structure coefficients for the variables. The criterion variables that most contributed to the discriminating power of Function I were: (a) Chances R will graduate high school ("What are the chances that you will graduate from high school?"); standardized function coefficient = .748; structure coefficient = .832; squared structure coefficient = .693; and (b) Number of friends dropped out ("How many of your friends have dropped out of school?"); standardized function coefficient = .224; structure coefficient = .396; squared structure coefficient = .157. Two other variables had relatively small standardized function coefficients but relatively large structure coefficients, indicating that these variables contributed to the underlying structure of Function I. These two variables were (a) R expects <= high school ("How far in school do you expect to go--beyond high school or not?"); standardized function coefficient =.12684; structure coefficient = .35409, and (b) number of friends to attend 4 yr college ("How many of your friends plan to attend a four year college?"); standardized function coefficient = .09818; structure coefficient = .30503.

Table 1
DDA Standardized Function and Structure Coefficients

Variable Label	Structure Coefficients I	II	III	Function Coefficients I	II	III
Chances R will graduate high school	**.832**	**-.142**	**-.087**	**.748**	**-.138**	**-.141**
How important is a good education	.093	-.045	.130	-.094	.041	.135
R expects <= to high school	**.354**	**-.535**	**-.022**	**.127**	**-.555**	**-.079**
R expects to graduate high school	.205	.154	.110	.058	.145	.010
R expects <= to college	.293	-.032	-.139	.033	.080	-.117
Every time I get ahead something stops me	.134	.067	-.014	.005	.045	.047
My plans hardly ever work out	.122	.063	-.086	.007	.123	-.157
I am satisfied with myself	.085	-.005	.013	.060	.037	.017
I don't have much to be proud of	.085	-.063	-.061	-.030	-.085	-.109
I don't have enough control over my life	.085	-.014	-.019	.016	.006	-.059
I'm a person of worth	.053	-.043	-.020	-.007	.066	-.185
When I make plans I can make them work	.041	-.004	-.002	.008	.007	.016
Chance and luck are important in my life	.109	-.134	.079	-.017	-.171	-.002
Good luck is more important than hard work	.067	-.094	.231	-.023	-.059	.272
I feel useless at times	.020	.001	-.112	-.008	.016	-.242
I am able to do things as well as others	.018	-.077	.077	-.045	-.068	.173
I feel good about myself	.020	-.058	.067	.023	-.051	.138
I think I'm no good at all	.021	-.011	.029	-.050	.042	.226
Does Respondent smoke	.135	.056	.008	.094	-.039	.003
Last 30 days, n of times drank alcohol: 0	-.023	.068	.005	-.065	-.039	.041
Last 30 days, n of times drank alcohol: > 2	-.005	.051	-.001	-.033	-.081	-.127
Last 30 days, n of times drank alcohol: > 19	.041	-.049	.190	-.016	-.132	.186
Last 12 months, n of times drank alcohol: 0	-.037	.134	-.049	-.043	.106	-.043
Last 12 months, n of times drank alcohol: > 2	-.017	.115	-.011	.051	.019	-.015
Last 12 months, n of times drank alcohol: > 19	.012	.119	.054	-.004	.070	.034
Ever used marijuana	.100	.194	-.041	-.006	.030	-.058
Ever used cocaine	.093	.199	.100	.014	.096	.089
Times late for school: 0	.044	.099	-.150	-.029	-.006	-.121
Times late for school in the last 4 weeks: >= 2	.089	.165	-.116	.028	.015	-.110
Times late for school in the last 4 weeks: >= 3	-.099	-.186	.040	.016	-.042	-.004
Times skipped classes: 0	.107	.162	.026	-.008	-.019	.163
Times skipped classes > 2	.170	.252	-.086	.015	.187	-.103
Times skipped classes > 3	-.197	-.195	.123	-.094	.009	.118
Number of times suspended	.276	.153	.099	.128	.054	.044
Times in-school suspension	.252	.130	.082	.041	.027	.033
Times in trouble for not following rules: 0	.113	.149	.083	-.045	-.034	.074
Times in trouble for not following rules: > 2	.173	.286	.118	.023	.185	-.086
Times in trouble for not following rules: >= 3	-.176	-.242	-.285	-.014	-.071	-.346
Ever been arrested	.124	.183	-.004	.063	.086	-.054
Does R belong to a gang	.045	.132	.029	-.004	.118	.032
How often spend time in religious activities	.087	-.068	-.008	.031	-.071	.086
Considers self religious	.063	-.035	-.071	-.003	-.055	-.133
Volunteer service	-.026	.096	-.085	-.027	.078	-.113
Community service	.075	-.091	-.078	.011	-.096	-.055

(table continues)

BEST COPY AVAILABLE

Table 1 (continued)

Variable Label	Structure Coefficients			Function Coefficients		
	I	II	III	I	II	III
Participated in school sports	-.016	-.213	-.029	-.028	-.167	-.043
Participated in sports lesson	-.008	-.086	-.074	-.020	-.025	-.076
Participated in any hobbies	-.008	-.118	.006	-.023	-.032	-.028
How often do you drive or ride around in a car	-.017	-.017	.147	.004	.097	.153
How often do you talk with other adults	-.084	-.119	.088	-.032	-.058	.135
R reports mom expects < high school	.282	-.332	.098	.091	-.093	.094
R reports mom expects >= to college	.271	-.205	-.045	.039	-.169	.071
R reports dad expects >= college	.248	-.064	-.118	-.022	.206	-.333
R reports dad expects > high school	.238	-.147	.025	.039	-.103	.229
Who decides how late R can stay out	.248	.189	.084	.111	.132	.070
Who decides if R can work	.177	.172	.021	.055	.098	.043
Who decides how R spends money	.095	.078	.038	.043	-.003	.055
How often do you talk/spend time w/parents	-.005	.002	-.069	-.121	-.026	-.176
Number of times family moved since 1988	.166	.207	.087	.060	.058	.286
Number of times changed schools since 1988	.179	.401	-.191	.124	.316	-.255
Moved in the last 2 years	-.076	-.125	.034	.031	.007	.110
Last 2 yrs, family member seriously ill	-.013	-.080	.207	.012	-.039	.158
Last 2 yrs, parent died	-.041	.056	.160	.007	.070	.156
Last 2 yrs parents divorced	-.051	-.055	.097	-.010	-.024	.085
Last 2 yrs, parent lost job	-.030	-.066	.045	.005	-.010	.022
Last 2 yrs, sibling drop out of school	-.131	-.098	-.189	-.036	-.095	-.253
last 2 yrs, family member in rehabilitation	-.032	-.061	.129	.031	.015	.113
Last 2 yrs, family member crime victim	-.016	-.071	.101	.019	.016	.080
Last 2 yrs, R seriously ill	-.017	.021	.019	.024	.064	-.041
Number of friends dropped out	**.396**	**.198**	**.187**	**.224**	**.113**	**.314**
Number of friends plan work full-time	.192	.101	-.207	-.027	.084	-.203
Number of friends with no plans for college	.151	.018	.019	-.008	-.070	.062
Number of friends to attend 2 year college	.027	.049	-.181	-.091	-.071	-.094
Number of friends to attend 4 yr college	**.305**	**.139**	**.042**	**.098**	**.109**	**.168**
Friends: important education past high school	.146	.089	-.049	.009	.025	-.233
Friends: important to attend classes	.118	.053	.056	.036	-.111	-.062
Friends: important to get good grades	.044	.081	.168	-.020	-.009	.222
Friends: important to study	.050	.085	.109	-.053	.003	.086
Friends: important to have a steady job	-.151	-.048	.159	-.060	-.040	.112
Friends: important to participate in religion	-.010	.072	-.078	-.022	.030	-.067
Friends: important to play sports	-.019	-.013	.048	-.065	.003	.078
Friends: important to use alcohol	-.022	.083	.081	-.056	-.005	.079
Friends: important to use drugs	.034	.114	.091	-.054	.036	.025
Number of friends in gangs	.103	.112	-.076	-.097	-.083	-.113
Many gangs in school	.136	.119	-.050	.035	.051	-.048
Fights occurred between racial/ethnic groups	.115	.115	.015	.009	.090	.031
Student disruptions a problem at school	.071	.036	-.030	-.003	.012	-.026
Students friends with other racial groups	.026	-.072	.037	.007	-.062	-.006
R did not feel safe at school	.091	-.004	.042	-.024	-.074	.078
There was real school spirit	-.007	-.076	-.032	-.021	-.096	-.069
Teachers interested in students	.092	-.035	-.037	.013	-.121	-.104
The teaching was good	.071	.048	.061	-.011	.115	.137

Note. Coefficients deemed to contribute the most to the discriminating power of Function I are **bolded.** R = Respondent.

The group centroids obtained in this DDA indicated that Function I primarily discriminated the stayin group (group 3) from the other three groups (groups 0, 1, 2). The stayins (M = +3.31653) and the early dropouts (M = -4.96028) differed most on this function as regards the response variables. The group centroids for this DDA are presented in Table 2.

Table 2
DDA Discriminant Function Group Centroids

Group	Function I	Function II	Function III
0	-3.04418	-1.07589	-0.14612
1	-4.96028	0.85244	-1.06731
2	-3.81855	0.31922	0.50211
3	3.31653	0.00923	-0.00289

Note. 0=alternative completer; 1=early dropout; 2=late dropout; 3=stayin.

Conclusion

A descriptive discriminant analysis was conducted to determine the effects of group membership in 1992 (i.e., alternative completer, early dropout, late dropout, stayin, or stayin vs dropouts) upon the response variables (data collected in 1992). Only one function was deemed meaningful in explaining the group differences. The underlying structure of this function was determined by assessing the relative magnitude of the standardized function coefficients and the structure coefficients for the response variables. These coefficients were presented in Table 1. This function primarily involved response variables regarding persons' assessment of (a) "their subjective chances of graduating from high school" (standardized function coefficient = .74846; structure coefficient = .69274; squared structure coefficient = .47989), (b) "their expectations for future education" (standardized function coefficient = .12684; structure coefficient = .35409; squared structure coefficient = .12538), (c) "their friends' expectations for future education" (standardized function coefficient = .12684; structure coefficient = .30503; squared structure coefficient = .09304), and (d) "the number of friends who had dropped out of high school" (standardized function coefficient = .22396; structure coefficient = .39592; squared structure coefficient = .15675). Therefore the underlying structure of this function primarily involved individual and peer educational expectations.

This is not a surprising finding considering the temporal proximity to the time of expected graduation for the members of the 1988 eighth-grade cohort. These findings do support previous research that high school dropouts are accurate assessors of their academic situation (Bachman et al., 1971; Peng & Takai, 1983; Rumberger, 1981, 1983). These findings are also consistent with previous research regarding peer influence on adolescent behavior (Gavin & Furman, 1989; Jessor & Jessor, 1977). That is, school dropouts tend to have friends who are dropouts and peer groups tend to share similar educational aspirations (Ekstrom et al., 1986; Pallas, 1985; Rumberger, 1987, 1995).

It is important to note that many response variables found in previous research to be strongly related to dropping out of high school did not prove to contribute to the discriminating power of the functions in the present study (e.g., variables related to parental involvement, school climate, social adjustment, delinquency (including alcohol and other drug use), family composition, and family transitions). However, these variables still might be useful for some non-descriptive or predictive purposes not considered here (Campbell, 1997). For example, the locus of control scale (Rotter, 1966) did not contribute to the discriminating power of the functions. This makes sense in that those students who drop out of school are likely to be leaving an environment in which they are failing, and in turn are gaining a sense of control in their lives. Those students who stay in school are about to attain a goal that they have worked toward for a long time (i.e., graduate from high school) and are also likely to perceive a heightened sense of control in their lives (Rumberger, 1983; Wehlage & Rutter, 1986). Thus both groups may experience a heightened sense of control. The lack of discriminating power of the locus of control scale may also be because having external locus of control makes a student more vulnerable to external pressures regardless of how the pressure is being applied. External pressures such as SES or peer pressure may influence a student to drop out of school (Bachman et al., 1971; Ekstrom et al., 1986; Rumberger, 1987), but external pressures such as parental attitudes toward education or peer attitudes toward education can also pressure a student to remain in school.

The self-concept scale also did not contribute to the discriminating power of the functions. This supports Marsh's (1994) "Big-Fish-Little-Pond-Effect." This effect occurs when students (or dropouts) compare their abilities to others in their reference group and self-concept is judged relative to their reference group. This results in dropouts comparing themselves to other dropouts and stayins comparing themselves to other stayins thus limiting discrimination between the two groups.

The consequences of dropping out of high school continue to become more severe both for the individual and for society. High school dropouts are likely to experience detriments in psychological functioning and negative economic effects at the personal level. Individuals who lack the skills and ability to compete in a high-tech, global market are destined to low-paying jobs or dependence upon family and/or society. The burden upon society to subsidize high school dropouts in terms of lower worker productivity, increased welfare and unemployment benefits, higher health care costs, and increased costs in the criminal justice system will continue to grow. Dropping out of high school is a complex process involving many systems (i.e., individual, family, school, community). More research aimed at understanding the multivariate, systemic dynamics involved in dropping out of school is needed in order to develop more effective prevention and intervention programs for students at-risk for dropping out of school. The present study employed national, longitudinal, large-sample data to understand better the dynamics of dropping out of high school.

References

Bachman, J. G., Green, S., & Wirtanen, I. D. (1971). *Youth in transition (Volume III): Dropping out - problem or symptom?* Ann Arbor, MI: Institute for Social Research. (ERIC Document Reproduction Service No. ED 059 333)

Baker, D. P., & Stevenson, D. L. (1986). Mother's strategies for children's school achievement: Managing the transition to high school. *Sociology of Education, 59*, 156-167.

Barro, S. M., & Kolstad, A. (1987). *Who drops out of high school? findings from High School and Beyond.* Washington, DC: U.S. Government Printing Office.

Bryk, A. S., & Thum, Y. M. (1989). The effect of high school organization on dropping out: An exploratory investigation. *American Educational Research Journal, 26*, 353-383.

Campbell, T. C. (1997). *Predicting educational status: A multivariate, multi-perspective, longitudinal analysis of the process of dropping out of secondary school.* Unpublished doctoral dissertation, Texas A&M University.

Catterall, J. S. (1985). *On the social costs of dropping out of school* (Report No. SEPI-86-3). Stanford, CA: Stanford Education Policy Institute. (ERIC Document Reproduction Service No. ED 271 837)

Catterall, J. S. (1986). *A process model of dropping out of school: Implications for research and policy in an era of raised academic standards* (Report No. OERI-G-86-0003). Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED 281 137)

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cooley, C. H. (1902). *Human nature and the social order.* New York: Scribner's.

Deschamps, A. B. (1993). An integrative review of research on characteristics of dropouts. (Doctoral dissertation, the George Washington University, 1992). *Dissertation Abstracts International, 54* (01), 137.

Ekstrom, R. B., Goertz, M. E., Pollack, J. M., & Rock, D. A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers College Record, 87*, 356-373.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59*, 117-142.

Finn, J. D., & Owings, M. F. (1994). Family structure and school performance in eighth grade. *The Journal of Research and Development in Education, 27*, 176-187.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Franklin, C., McNeil, J. S., & Wright, R. (1990). School social work works: Findings from an alternative school for dropouts. *Social Work in Education, 12*, 177-194.

Franklin, C., & Streeter, J. S. (1990, October). *Differentiating characteristics between high achieving/high income and low achieving/low income drop out youth: Considerations for treatment programs.* Paper presented at the Fourteenth Annual National Association of Social Workers Texas State Convention, El Paso, TX.

Frase, M. J. (1989). *Dropout rates in the United States: 1988 Analysis Report* (NCES 89-609). Washington, DC: U.S. Department of Education. (ERIC Document Reproduction Service No. ED 313 947)

Gavin, L. A., & Furman, W. (1989). Age differences in adolescents' perceptions of their peer groups. *Developmental Psychology, 25*, 827-834.

Haggerty, C., Dugoni, B., Reed, L., Cederlund, A., & Taylor, J. (1996). *National Educational Study (NELS:88/94) methodology report* (NCES 96-174). Washington, DC: U.S. Department of Education

Higgins, K. B. (1987). *Adolescent self-esteem and the schools.* (ERIC Document Reproduction Service No. ED 312 528)

Huberty, C. J. (1994). *Applied discriminant analysis.* New York: John Wiley & Sons.

Huberty, C. J., & Barton, R. M. (1989). An introduction to discriminant analysis. *Measurement and Evaluation in Counseling and Development, 22*, 158-168.

Ingels, S. J., Abraham, S. Y., Karr, R., Spencer, B. D., Frankel, M. R., & Owings, J. A. (1990). *National Educational Longitudinal Study of 1988. Base year:*

Student component data file user's manual (NCES 90-464). Washington, DC: U.S. Department of Education.

Ingels, S. J., Dowd, K. L., Stipe, J. L., Baldridge, J. D., Bartot, V. H., Frankel, M. R., & Quinn, P. (1994). *National Educational Longitudinal Study of 1988. Second follow-up: Dropout component data file user's manual* (NCES 94-375). Washington, DC: U.S. Department of Education.

Ingels, S. J., Dowd, K. L., Baldridge, J. D., Stipe, J. L., Bartot, V. H., Frankel, M. R., & Quinn, P. (1994). *National Educational Longitudinal Study of 1988. Second follow-up: Student component data file user's manual* (NCES 94-374). Washington, DC: U.S. Department of Education.

Jessor, R., & Jessor, S. L. (1977). *Problem behavior and psychosocial development: A longitudinal study of youth.* New York: Academic Press.

Kaplan, D. S., Damphousse, K. R., & Kaplan, H. B. (1994). Mental health implications of not graduating from high school. *Journal of Experimental Education, 62*, 105-123.

Kaufman, P., Bradby, D., & Owings, J. (1992). *Characteristics of at-risk students in NELS:88* (NCES 92-042). Washington, DC: U.S. Department of Education. (ERIC Document Reproduction Service No. ED 349 369)

Klecka, W. R. (1980). *Discriminant analysis.* Beverly Hills, CA: Sage.

Mare, R. D. (1980). Social background and school continuation decisions. *Journal of the American Statistical Association, 75*, 295-305.

Marsh, H. W. (1994). Using the National Longitudinal Study of 1988 to evaluate theoretical models of self-concept: The Self-Description Questionnaire. *Journal of Educational Psychology, 86*, 439-456.

McDill, E. L., Natriello, G., & Pallas, A. (1987). The high costs of high standards: School reform and dropouts. In W. T. Denton (Ed.), *Dropouts, pushouts, and other casualties* (pp. 183-209). Bloomington, IN: Phi Delta Kappa.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition.* New York: John Wiley & Sons.

McMillen, M. M. (1992). *Eighth to tenth grade dropouts. Statistics in brief* (NCES 92-006). Washington, DC: U.S. Department of Education. (ERIC Document Reproduction Service No. ED 342 871)

Mensch, B. S., & Kandel, D. B. (1988). Dropping out of high school and drug involvement. *Sociology of Education, 61*, 95-113.

Pallas, A. M. (1987). School dropouts in the United States. In W. T. Denton (Ed.), *Dropouts, pushouts and other casualties* (pp. 23-39). Bloomington, IN: Phi Delta Kappa. (Reprinted from *The condition of education,* (pp. 158-174), by J. D. Stern & M. F. Williams, Eds., 1986, Washington, DC: Center for Education Statistics)

Pallas, A. M. (1985). The determinants of high school dropout. *Dissertation Abstracts International, 45* (12), 3605. (University Microfilms No. AAC-8501672)

Pallas, A., Natriello, G., & McDill, E. (1989). The changing nature of the disadvantaged population: Current dimensions and future trends. *Educational Researcher, 18*, 16-22.

Peng, S. S., & Lee, R. M. (1992). *Measuring student at-riskness by demographic characteristics.* (ERIC Document Reproduction Service No. ED 347 679)

Peng, S. S., & Takai, R. T. (1983). *High school dropouts: descriptive information from High School and Beyond* (NCES 83-221-b). Washington, DC: U.S. Department of Education.

Rock, D. A., & Pollack, J. M. (1995). *Psychometric report for the NELS:88 base year through second follow-up* (NCES 95-382). Washington, DC: U.S. Department of Education.

Roderick, M. (1993). *The path to dropping out: Evidence for intervention.* Westport, CT: Auburn House.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80* (1, Whole No. 609).

Rumberger, R. W. (1981). *Why kids drop out of high school* (Program Report No. 81-B4). Stanford, CA: Institute for Research on Educational Finance and Governance.

Rumberger, R. W. (1983). Dropping out of high school: The influence of race, sex, and family background. *American Educational Research Journal, 20*, 199-220.

Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. *Review of Educational Research, 57*, 101-121.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*, 583-625.

Schulz, E. M., Toles, R., & Rice, W. (1987). The association of dropout rates with student attributes. In W. T. Denton (Ed.), *Dropouts, pushouts, and other casualties* (pp. 89-103). Bloomington, IN: Phi Delta Kappa.

Sewell, T. S., Palmo, A. J., & Manni, J. L. (1981). High school dropout: Psychological, academic, and vocational factors. *Urban Education, 16*, 65-76.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1994, February). *Why multivariate methods are usually vital in research: Some basic concepts*. Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Society for Research in Human Development (SWSHRD), Austin, TX. (ERIC Document Reproduction Service No. ED 367 687)

Thompson, B. (1995). Review of *Applied discriminant analysis* by C. J Huberty. *Educational and Psychological Measurement, 55*, 340-350

Thompson, B. (1997a). Editorial policies regarding statistical significance testing: Further comments. *Educational Researcher, 26*(5), 29-32.

Thompson, B. (1997b). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement, 57*, 5-19.

Tidwell, R. (1988). Dropouts speak out: Qualitative data on early school departures. *Adolescence, 23*, 92.

Voydanoff, P., & Donnelly, B. W. (1990). *Adolescent sexuality and pregnancy*. Newbury Park, CA: Sage.

Wehlage, G. G. (1987). At-risk students and the need for high school reform. In W. T. Denton (Ed.), *Dropouts, pushouts, and other casualties* (pp. 211-221). Bloomington, IN: Phi Delta Kappa.

Wehlage, G. G. (1989). Dropping out: Can schools be expected to prevent it? In L. Weis, E. Farrar, & H. G. Petrie (Eds.), *Dropouts from school: Issues, dilemmas, and solutions* (pp. 1-19). Albany, NY: State University of New York Press.

Wehlage, G. G., & Rutter, R. A. (1986). Dropping out: How much do schools contribute to the problem? *Teachers College Record, 87*, 374-392.

Wehlage, G. G., Rutter, R. A., & Turnbaugh, A. (1987). A program model for at-risk high school students. In W. T. Denton (Ed.), *Dropouts pushouts, and other casualties* (pp. 155-169). Bloomington, IN: Phi Delta Kappa.

Weis, L., Farrar, E., & Petrie, H. G. (Eds.). (1989). *Dropouts from school*. Albany, NY: State University of New York Press

Weishew, N. L., & Peng, S. S. (1993). Variables predicting students' problem behaviors. *Journal of Educational Research, 87*, 5-17.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 11-17

Applications, Trials, and Successes of CU-SeeMe in K-12 Classrooms in the Southeast

Anna C. McFadden, Margaret L. Rice, Vivian H. Wright
The University of Alabama

Roger Saphore
Alabama State University

Jenka Keizer
The University of Alabama

Desktop videoconferencing with CU-SeeMe offers educators an alternative for enhancing learning in K-12 classrooms. In an educational setting, CU-SeeMe can offer a connection with external resources; support the use of diverse media; and enable document sharing, facilitating collaboration and feedback. This study examined whether CU-SeeMe is being utilized in classrooms and if so, how. A survey was sent to registered CU-SeeMe users in Alabama, Arkansas, Georgia, Kentucky, Louisiana, Mississippi, and Tennessee via electronic mail. Results indicated that users participating in the study believed that CU-SeeMe enriched the learning environment in their classrooms. They stated many reasons for this belief, and provided examples of activities in which students have been involved. When conducting CU-SeeMe activities, the students are the active participants, initiating and setting up conferences, designing and solving problems, developing questions to elicit information on particular topics, conducting interviews, handling audio transmissions or running the keyboard, and conducting troubleshooting. The teacher acts as an observer and facilitator.

As technological developments continue, educators continue to attempt to integrate these new developments into the learning process. One of the current technologies being integrated into some K-12 classrooms is videoconferencing. Videoconferencing allows simultaneous visual and oral communication in real-time with two or more individuals from different locations. Participants can communicate and share computer documents and resources while on-line. A videoconferencing system requires audio and visual equipment, generally consisting of a monitor, camera, microphone, and speakers connected to a computer. Options for communicating between sites include modems, satellites, and ISDN lines.

There are many videoconferencing systems available, and expense varies with the type of system. CU-SeeMe is a videoconferencing software program gaining popularity with educators because it is less expensive than other videoconferencing software and, therefore, more accessible to K-12 schools. CU-SeeMe users also may download a copy for a 30-day, limited use trial. Many educators select CU-SeeMe for videoconferencing because of the trial use and the low purchase price. With CU-SeeMe, if a user wishes to just view video sent by another user, and not send video, a camera is not necessary. However, sending video to another user does require either a special type of video camera with a parallel port hookup or a video capture board and a regular, home video camera.

Videoconferencing is still in its infancy stage in schools and limited research could be located regarding videoconferencing in K-12 schools (Edmonds, 1996; Roblyer, 1997; Todd, 1996). There has been little research conducted to investigate whether many K-12 schools are using CU-SeeMe, how they are using it, and to record various trials and successes experienced. This study examined whether CU-SeeMe software is being used in K-12 schools in the Southeast, and if so, how.

Statement of the Problem

In the recent past, videoconferencing technology was typically an expensive undertaking, one most K-12 schools could not afford. With the development of new technology, however, videoconferencing is now within

---

Anna C. McFadden is Associate Professor and Margaret L. Rice is Assistant Professor in Instructional Technology at The University of Alabama. Vivian H. Wright is a Doctoral Candidate in Instructional Technology at The University of Alabama. Roger Saphore is an Instructor of Library Media & Technology at the Alabama State University. Jenka Keizer is a Doctoral Candidate in Instructional Technology at The University of Alabama. Correspondence regarding this article should be addressed to Margaret L. Rice, The University of Alabama, Box 870203, 204 Wilson Hall, Tuscaloosa, Alabama 35487 or by e-mail to mrice@bamaed.ua.edu.

reach of many educational institutions. There are free or inexpensive programs, such as CU-SeeMe, available, and videoconferencing can be conducted on-line over the Internet. King (1996), in an on-line article, states:

> Although soothsayers have been predicting a videoconferencing boom for a number of years, industry experts say the pieces are in place to make it really happen this time. Those pieces include rising demand from multinational companies, improvements in technology, solidification of key standards, and proliferation of standards-compliant video-enabled products from heavy hitters like Microsoft Corp. and Intel Corp.

Many researchers discuss the advantages of videoconferencing for all aspects of our society. As a medium, videoconferencing is believed to decrease travel costs, enhance the written and spoken word, and facilitate shared software applications productivity (Czeck, 1995). Czeck, on-line, writes,

> Desktop videoconferencing has the potential to enhance verbal and written communication, as well as to increase the efficiency of communication for group work. Even though the medium has the possibility of enriching communication, a person could decide that it is not worth the problems it creates.

Fetterman (1996) discusses benefits directly related to education. Videoconferencing enhances collegial communications, and offers more "personal" electronic communication where nonverbal cues are available. Other benefits include instructional support, quick responses from remote locations, and crossing distance informational barriers (Littman, 1995). With CU-SeeMe, schools can attempt to keep up to date with the "real world." Hudson, on-line, (1996), comments, "Distance learning and counseling/mentoring benefit from one-to-one videoconferencing. . . . University professors or students advise K-12 students on career information or research projects" through the use of videoconferencing programs like CU-SeeMe. Duffy (1996), on-line, comments "We now have a virtual education reality. We all become students and teachers simultaneously. The term student-teacher takes on a whole new meaning!"

Reed and Woodruff (1997) list the following videoconferencing advantages for educators: Establishes a visual connection among participants; enables connection with external resources; supports use of diverse media; and enables document sharing, facilitating collaboration, and feedback. Reed and Woodruff further suggest that instructors make the shift from "knowledge disseminator" to "learning facilitator."

Czeck (1995) also discusses some disadvantages for videoconferencing: users experiencing discomfort toward the system and anxiety about the use of video; information overload; surveillance issues; and technological problems. Other disadvantages include: many affordable videoconferencing systems normally do not transmit a high quality video image; there is a lack of industry standards; and classroom teachers cannot control such technological issues. While teachers and students find technology relevant and useful, it is in a state of change. Dyrli and Kinnaman (1995) tell us that educators must realize that unlike anything we have known in the past, where content changed but the delivery format did not, technology is not static. The medium itself changes rapidly.

Although researchers agree on the benefits of videoconferencing, little research has been conducted on the actual use of videoconferencing in K-12 schools. Todd (1996) reported on the use of CU-SeeMe in a school in Virginia to communicate with international sites, including educational communities in New Zealand. Various activities were described. Edmonds (1996) described a study in Australia that found students believed they had better contact with teachers through videoconferencing. Additionally, the study found that medically disabled students who received instruction via videoconferencing exceeded expectations for improvement.

The current study addressed several research questions regarding the use of CU-SeeMe videoconferencing software in K-12 schools. Potential benefits of videoconferencing are evident, but are educators utilizing these opportunities, and if so, how? In an effort to determine CU-SeeMe usage in K-12 schools, the research team determined the following questions for exploration:

- Which schools are using CU-SeeMe software?
- For what primary purpose are schools using CU-SeeMe?
- Do users of CU-SeeMe find the software easy/difficult to use?
- What type of involvement do the students have when CU-SeeMe is used?
- What improvements are needed for future CU-SeeMe use in schools?

Methodology

Currently, teachers using CU-SeeMe are invited to register their school name, contact name, and electronic mail address with the Internet site of Global School Net (http://www.gsn.org). On-line research determined that this site lists most CU-SeeMe educational users from the United States and other countries who choose to register

CU-SeeMe use. There may be other users who choose not to register.

This study surveyed the population registered as CU-SeeMe users in the southeastern states of Alabama, Arkansas, Georgia, Kentucky, Louisiana, Mississippi, and Tennessee. Users in the Southeast were selected for the study because the researchers are located in a southeastern state and were interested in this area for the initial study. Also, since the study was being conducting using email, it was believed that the sample size of 53 users in the Southeast would be manageable. Future research will include surveying other areas of the United States and international locations.

A survey with 20 Likert-type items (Appendix A), 10 open-ended questions, and demographic questions was developed for data collection via electronic mail. The survey instructions stated that survey completion was voluntary and that since the method of response was through electronic mail, the respondent was known. However, the participants were informed that the data collected would be reported as cumulative, not individual, and data and responses would be reported anonymously. Any reference to a specific response did not link or identify the respondent or the respondent's school. A pilot study was conducted with users in the state of Florida. No significant changes to the survey instrument were determined from the results of the pilot study.

Data Analysis and Results

During the week of April 29, 1997, the initial survey was sent to 53 e-mail addresses in the Southeast registered as CU-SeeMe users with the Global School Net Internet site. A follow-up message was sent during the week of May 20, 1997. A total of 18 individuals either responded to the initial e-mail message or completed the survey.

A return response was integrated into the heading for the follow-up message. There were two reasons for this action, first to track that the messages were read by someone and second, to demonstrate the importance of the survey process. Of the 53 surveys administered by e-mail, seven were completed and returned. Five were in response to the follow-up message. The researchers believe possible reasons for the low response were:

- Some of the respondents indicated that they were too busy. To avoid end of school year conflicts, the surveys were sent in late April. In retrospect, this time period probably conflicted with many Spring break dates.
- E-mail messages are overlooked. In the "big picture" of the typical school day, attention focuses on many critical areas. Perhaps this will change, as the overall population becomes more involved with technology and e-mail usage.
- The school is not using the CU-SeeMe package. Either it may be too complicated or it may be intimidating for some users.
- The CU-SeeMe users do not want to admit non-use of purchased school budget items.
- A possible reason for the non-response to the survey may be apathy on the part of the user. Perhaps the user was not involved in the initial equipment selection and purchase.

The low response rate limits the generalizability of the results of this study to other areas or to the non-respondents to the survey. This is a serious limitation of the study, however, the responses do provide valuable information for other individuals considering using videoconferencing in the learning process.

Demographic Data

Three of the seven respondents were from the same state. All respondents work at public schools with student populations from 500 to 1800 students. Class sizes average 27 students. The respondents' communities are predominately middle class areas. Respondents included two males and five females. Four respondents work at high schools, one at a junior high school and one at an elementary school (one is a district administrator). Four educators are computer lab instructors, and two are language teachers.

CU-SeeMe Use

CU-SeeMe use ranged from the novice level (1 to 3 months), to the intermediate (4 to 6 months), and expert levels (over 12 months.) Hardware was split between lab use and classroom use. Respondents had installed both software and hardware and had experienced minor difficulties with installation. All respondents strongly agreed that using CU-SeeMe was fun. They were split on the ease of use. Three respondents strongly agreed that using CU-SeeMe was easy, two were undecided, and two disagreed. Two respondents also agreed that they found CU-SeeMe frustrating to use, while two respondents were undecided, and three disagreed. One user advised that new users of CU-SeeMe should be careful not to become frustrated with their first few connections and should continue trying. All respondents agreed that CU-SeeMe enriches their classroom learning experiences.

When asked to indicate with whom CU-SeeMe is used to communicate, only one respondent indicated that it is used to communicate with other classrooms in their own school. Two respondents indicated that it is used to communicate with other schools in their system, and six noted communication with other schools in the United States. Four of the respondents use CU-SeeMe to communicate with students in other countries. None of the respondents used CU-SeeMe to communicate with members of their community. Four responded that they use CU-SeeMe to communicate with experts from various fields, three communicate with government officials, and four use it to communicate with school administrators.

CU-SeeMe Activities

Three respondents use CU-SeeMe to take electronic field trips to commercial sites such as CNN, and educational sites such as NASA. One respondent indicated using CU-SeeMe for field trips to government sites. Four of the respondents use CU-SeeMe in large (10 or more students) and small (under 10 students) group activities and for individual activities. Group sizes ranged from three students to whole classes.

When asked about the types of activities for which CU-SeeMe is used, a variety of responses was received. One respondent stated that CU-SeeMe is used for data collection, cultural exploration, interviews, to share information, for school contests, and to conduct instruction via conferencing. Other responses included sharing ideas with other classrooms and students, conferencing with experts to enhance classroom units, and communicating across town with the central office. One respondent's class uses it to communicate with Spanish classes with whom they also communicate through e-mail. Another group, which consists of new users, has used it in science classes and for holiday activities. The primary activities were listed as collaboration, enhancing the curriculum, alternative communication, and communicating with other classrooms.

The respondents indicated that their students take active roles during CU-SeeMe sessions. Students are expected to initiate and set up conferences, act as facilitators, design and solve problems, design questions to elicit information on particular topics, conduct interviews, handle audio transmissions or run the keyboard, and troubleshoot. Generally the teacher acts as an observer or facilitator.

Problems and Needed Improvements

Problems encountered with CU-SeeMe included installation of the hardware and software, difficulty in finding appropriate projects, sparse documentation, unwanted participants if a conference is not scheduled, time zone differences, and the expense of running CU-SeeMe on every classroom computer. When asked what types of problems they encountered when installing the software, those that had problems indicated problems with the software crashing and destroying the selected preferences and with the hostname. Problems with hardware included difficulty in using a video camera other than QuickCam and technical problems with audio. A few respondents had difficulty setting up conferences with schools in other time zones or missing conferences in which they wished to participate due to the time differences. An area of concern to the respondents was the lack of control over inappropriate and indecent materials. Unauthorized individuals breaking into the videoconferencing sessions and broadcasting indecency to the students are perceived as hindrances to CU-SeeMe use. One respondent suggested that if more schools participated, the educational environment could possibly initiate more control.

Areas where respondents indicated improvement is needed include the difficulty in finding IP (Internet Protocol) addresses and projects on the Internet, poor audio quality, time differences for other areas, people not online when they are scheduled, and a lack of details regarding the subject area that will be covered during the conference. One respondent suggested that some type of handbook would be helpful during installation.

Advantages and Benefits of Using CU-SeeMe

The respondents to the survey saw many definite advantages to using CU-SeeMe in the classroom. One respondent wrote,

> Students have the opportunity to 'see' the people that they are communicating with; makes the exercise more personal; our biology teachers have connected to other classrooms who have their computers connected to large screen TVs or LCD panels and conducted experiments themselves or had their students conduct experiments. Students in the other classrooms then ask questions and get immediate responses.

Another respondent saw as advantages the enthusiasm of the students; the experience and knowledge that were gained by the students; and the bringing together of students with people who could enrich their learning, people with whom the students otherwise would have no opportunity to converse. The respondent who used it to communicate with Spanish classes felt that it provided students with another method to enhance their language usage. According to Fetterman (1996), "electronic communication is a little more personal and a lot more effective when you can hear the nuances of tone and see nonverbal 'language' such as gestures and expressions. . . ." (p. 23). One respondent wrote that "The students really enjoy being able to see and talk to students across the US. The ability to weigh problems amongst classes allows students to look at material presented in many different lights."

Benefits to having more than one CU-SeeMe connection in a classroom were that conferences could be conducted with more than one school or individual at a time, which would decrease the amount of class time needed for conferencing. "Because of physical configuration, video and audio, it would be more effective to have three or four stations in a lab or larger classroom." There would be more use of the videoconferencing and less time spent waiting. More than one class at a time could participate: "With more stations within our building, classes would be able to locate their own CU-SeeMe conferences and participate without me being present."

Conclusions and Discussion

The most important finding of this study was that, although the sample was very small (i.e., seven participants), all of the participants believe that CU-SeeMe enriches their classroom learning environment. The participants stated many reasons why they believe this, and provided examples of the types of activities in which the students have been involved. Some of these activities included data collection, cultural exploration, interviews, idea and information sharing, school contests, conducting instruction, conferencing with experts to enhance classroom units, and communicating across town with the central office. One respondent's class uses it to communicate with Spanish classes with whom they also communicate through e-mail. The primary activities were listed as collaboration, enhancing the curriculum, alternative communication, and communication with other classrooms. Although the sample for this study was small, the data demonstrate that CU-SeeMe is being used in K-12 schools. Additional studies now need to be conducted to examine its effectiveness in enhancing the learning process.

Many of the current education reform efforts encourage the construction of knowledge, where students are active participants and teachers often act as facilitators. Data from this study suggest that CU-SeeMe can aid in these reform efforts. The participants in this study indicated that when conducting CU-SeeMe activities, the students are the active participants, initiating and setting up conferences, designing and solving problems, developing questions to elicit information on particular topics, conducting interviews, handling audio transmissions or running the keyboard, and conducting troubleshooting. The teacher acts as an observer and facilitator.

Users reported that they experienced minor difficulties with installation of hardware and software, but overall they agreed that CU-SeeMe is fun and easy to use. One user suggested that a handbook would be helpful. Cornell University has a website (http://cu-seeme.cornell.edu/) that contains information about CU-SeeMe. There is a user's guide, a listserv (http://cu-seeme.cornell.edu/listinfo.html), and information on licensing and copyright. There are also many websites available that contain general information about videoconferencing. The Pacific Bell website (http://www.kn.pacbell.com/wired/vidconf/) contains a wealth of information about videoconferencing projects, technical information, newsgroups, and e-mail lists. This website also contains materials for instructor and participant evaluation of videoconferences.

Although videoconferencing is still in its infancy in schools, CU-SeeMe's perceived ease of use and educational flexibility make it a viable tool for the school of the future. Videoconferencing allows more interaction with other users than e-mail or Internet, and can be used to enhance learning, especially in areas such as learning other languages or about other cultures. The literature review conducted for this study and the responses from the participants identified videoconferencing's positive counseling and mentoring potential with experts and professionals around the world from diverse cultures and work place settings. Some educators refer to this type of learning experience as electronic field trips. Such trips can aid in personal communication and can cross distance informational barriers at minimal costs while sharing ideas and documents in collaborative efforts. Additionally, such learning experiences can decrease the amount of class time needed for similar collaboration.

Although the research is sparse on the effectiveness of videoconferencing, there are many Internet sites that describe projects currently being conducted in schools using CU-SeeMe. One such site is maintained by Rose City Park School in North East Portland, Oregon (http:www.teleport.com/~rcplib). One of the fifth grade classes at Rose City Park School has used CU-SeeMe to engage in collaborative learning with a third grade class in San Antonio, Texas and share projects. The students have conferenced together and created HyperStudio stacks. They also used the Internet to keypal with students in South Africa and Sweden and used CU-SeeMe to conference with these keypals. If school personnel are considering the use of videoconferencing, they may wish to access some of these projects to see how other classrooms use the software.

As Cu-SeeMe and other videoconferencing software become more widely used in K-12 schools, research will need to be conducted concerning the effectiveness of using videoconferencing and its ability to enhance learning. Studies similar to the current one have been planned using larger and more diverse samples. Other issues that need to be explored for future CU-SeeMe usage include monitoring of sessions to avoid unwanted participants, time zones issues, access to educational sites, audio quality, and being able to locate IP addresses and projects. Educators using CU-SeeMe may want to check the Global School Network Internet site and contact other CU-SeeMe users.

References

Czeck, R. (1995). *Desktop videoconferencing: The benefits and disadvantages to communication.* [On-line]. Available: http://ils.unc.edu/~czecr/papers/cscwpaper.html

Dyrli, O., & Kinnaman, D. E. (1995). What Every Teacher Needs to Know about Technology. Part I: Technology in Education: Getting the Upper Hand. *Technology & Learning*, 15(4), 37-43.

Duffy, L. (1996). *Scientist-on-Tap.* [On-line]. Available: http://www.gsn.org/gsn/proj/sot.home.html

Edmonds, R. (1996). *Distance teaching with vision.* (ERIC Document Reproduction Service No. ED 396 724)

Fetterman, D. M. (1996). Videoconferencing on-line: Enhancing communication over the Internet. *Educational Researcher*, *25*(4), 23-27.

Hudson, R. (1996). *SUCCEED: Deliverable Team #5: Electronic connectivity home page*. [On-line]. Available: http://www.visc.vt.edu/succeed/index.html

King, R. (1996). *A work (Still) in progress.* [On-line]. Available: http://www.teledotcom.com/1296/features/tdc1296video.html

Littman, M. K. (1995). Videoconferencing as a communications enhancement. *Journal of Academic Librarianship, 21*(5): 359-364.

Reed, J. & Woodruff, M. (1997). *Using compressed video for distance learning.* [On-line]. Available: http://www.kn.pacbell.com/wired/vidconf/Using.html

Todd, S. (1996). Going global: Desktop video conferencing with CU-SeeMe. *Learning and Leading with Technology, 24*(10), 57-61.

Appendix A: Survey Instrument

DEMOGRAPHIC DATA

A. Do you work in a public or private school?

B. What is the approximate size of your school?

C. How many students are in your class?

D. Would you consider your community to be a high, middle, or low socioeconomic area?

E. Male/Female

F. Grade Level you teach:

G. Subject you teach:

H. Grade Levels that use CU-SeeMe:

I. I have been using CU-SeeMe in my school for 0-1 month; 1-3 months; 4-6 months; 7-12 months; over 12 months:

J. Where is the CU-SeeMe that you are using located (i.e., classroom, library,lab)?

K. Did you install the CU-SeeMe software?

L. If so, did you have problems installing the CU-SeeMe software?

M. If yes, describe the specific problems and how you were able solve them.

N. Did you install the CU-SeeMe hardware?

O. Did you have problems installing the hardware necessary for using CU-SeeMe?

P. Describe the specific problems and how you were able to solve them.

(Additional Likert Scale items follow)

***************LIKERT SCALE******************

For the items 1-20, use the following scale and type in your answer directly following the item.

Scale= Strongly Disagree (SD), Disagree (D), Undecided (U), Agree (A), Strongly Agree (SA)

For example, if you strongly agree with a statement, type SA after the statement:

I find using the Internet for electronic field trips to be fun: SA

1. I find using CU-SeeMe to be fun:
2. I find using CU-SeeMe frustrating:
3. Using CU-SeeMe is easy:
4. Using CU-SeeMe is difficult:
5. I find using CU-SeeMe enriches my classroom's learning experiences:
6. We use CU-SeeMe to communicate with other classrooms in our school:
7. We use CU-SeeMe to communicate with other schools in our system:
8. We use CU-SeeMe to communicate with other schools in the United States:
9. We use CU-SeeMe to communicate with other students in other countries:

10. We use CU-SeeMe to communicate with people in our community:
11. We use CU-SeeMe to communicate with experts in various fields:
12. We use CU-SeeMe to communicate with government officials:
13. We use CU-SeeMe to communicate with school administrators:
14. We use CU-SeeMe for videoconferencing:
15. We use CU-SeeMe for fieldtrips to commercial sites such as CNN:
16. We use CU-SeeMe for fieldtrips to educational sites such as NASA:
17. We use CU-SeeMe for fieldtrips to government sites:
18. When we use CU-SeeMe, it is a large group activity (10 or more students):
19. When we use CU-SeeMe, it is a small group activity (less than 10 students):
20. When we use CU-SeeMe, it is an individual activity:

PLEASE TYPE ANSWERS TO THE FOLLOWING QUESTIONS:

AA. For what types of activities do you use CU-SeeMe in your school or classroom?

BB. What is the primary activity for which you use CU-SeeMe?

CC. When you use CU-SeeMe, how many of your students are involved in various activities?

DD. When you use CU-SeeMe for various activities, what are the students' roles?

EE. What suggestions do you have for improving CU-SeeMe for use in a school setting?

FF. What are the advantages of using CU-SeeMe in a school setting?

GG. What are the problems of using CU-SeeMe in a school setting?

HH. How would your class benefit with additional CU-SeeMe stations?

II. If other schools were interested, would you recommend the use of CU-SeeMe? Why or why not?

JJ. Does the use of CU-SeeMe augment your classroom's learning environment?

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 18-26

Secondary School Size and Achievement in Georgia Public Schools

Ellice P. Martin and John R. Slate
Valdosta State University

Relationships among secondary school size and academic achievement were investigated through an analysis of data obtained from the 1996 Georgia Public Education Report Card. Reading, math, and writing scores were analyzed to determine whether statistically significant differences were present as a function of secondary school size. Students in small schools performed more poorly on the Tests of Academic Proficiency (TAP) Math, TAP Reading, Georgia High School Graduation Test (GHSGT) Writing, GHSGT English, and GHSGT Math than did students in large schools. When socioeconomic factors were considered, school size was unrelated to student achievement. Implications and limitations are discussed.

Secondary school size was identified in the 1950s as an important factor in the effectiveness of a school (Conant, 1959). Since that time, research, theory, and practice have been directed toward determining the effect of school size on learning. Findings from studies and opinions from writings, especially about whether school consolidations accomplished the purpose of providing a better quality education, have been mixed (Fox, 1981; Franklin & Crone, 1992; McGuffey, 1991; Sher & Tompkins, 1977; Swanson, 1988; Walberg, 1992). Major questions related to school size and quality remain unanswered.

The factors believed to be reflective of school quality and to which school size may be related vary from study to study. For example, some researchers (e.g., Barker & Gump, 1964; Friedkin & Necochea, 1988; Howley, 1995) have studied relationships between school size and student achievement whereas other researchers (e.g., Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966; Forbes, Fortune, & Packard, 1993; McGuffey & Brown, 1978) have considered curriculum, climate, economic factors, and completion rates. Other factors have been identified as important influences of school size and of student achievement. Socioeconomic status (SES) of students is one such factor, one beyond the control of policy, but one which has been documented to influence student performance (e.g., Crone & Tashakkori, 1992; Friedkin & Necochea, 1988; Howley, 1995). These researchers, among others, have found evidence that student SES may mitigate or exacerbate the effects of school size on student achievement. Another factor, student dropout rate, has been linked to school size and student achievement (Costenbader & Markson, 1994; Pittman & Haughwout, 1987). Yet another variable, school expenditures, had been studied as an economic factor which can contribute to consolidation and larger schools (Burrup, Brimley, & Garfield, 1988; McGuffey & Brown, 1978; Walberg & Walberg, 1994).

An obvious assumption of citizens and decision makers in school consolidation is that a new building, centrally located for all of the children of a county, with all new equipment and the most modern curriculum, would provide a better education and would provide it more economically then several smaller facilities. In studies where researchers have examined the issue of school size and student achievement, Fox (1981) reported the presence of confounding variables. That is, in some studies, school district level data rather than individual school data were used. In other studies, ones in which population densities were considered, results were in conflict with studies in which population densities were not considered. In yet other studies, Fox (1981) stated that school size and student achievement results differed as a function of rural or urban setting.

Results from studies of school size and student achievement can, at best, be said to be conflicting. Even with the conflicting results, achievement does appear as a factor in most studies of school size. The definitions of small schools and large schools were, of course, relative to most individual studies, and that fact is stipulated. Results are reported here based on the terms used in a study by its authors.

Ellice P. Martin is a doctoral candidate in the Curriculum and Instruction track of the Ed.D. program at Valdosta State University. She is currently a secondary mathematics teacher at Clinch County High School in Homerville, GA. John R. Slate, Ph.D. is a Professor in the Department of Educational Leadership at Valdosta State University. Correspondence regarding this article should be directed to: Ellice P. Martin, Clinch County High School, 1011 Carswell Street, Homerville, GA 31634 or e-mail epmartin@planttel.net

Cotton (1997) considered 103 documents about school size, 31 of which addressed school size and achievement. About half of these 31 documents that addressed achievement were studies that resulted in no finding of significant difference between achievement levels of large and small schools. Authors of the other half concluded that student academic achievement in small schools was superior to that in large schools. None of the studies considered by Cotton found large schools superior to small schools in their effects on student academic achievement. Studies that fit in each of these categories are discussed below.

Beginning with an early study (Coleman et al., 1966), some researchers concluded that school size had no significant effect on achievement, either statistically or practically. One caveat, however, was that many authors reached that conclusion after controlling for some social or economic variable(s). Coleman et al. concluded that, once social composition of a school population was controlled, school characteristics did not significantly affect achievement. Stecklenberg (1991) sttudied Georgia student academic achievement and school size specifically. Though the correlation was slightly positive when socioeconomic factors were controlled, the author found little practical correlation between school size and the statewide tests in reading and math.

Other authors of studies prior to 1990 stated that school size was not a truly definitive factor in school quality as evidenced by student academic achievement. Harnisch (1987) found a statistically significant, though small, positive relationship between achievement and school size. The school size correlation (.13) accounted for only about 2% of the variance in achievement. Forbes, Fortune, and Packard (1993) found that achievement in some academic subjects varied across size levels, and achievement in other academic subjects was independent of size.

Recent studies that did not find school size statistically significantly related to student academic achievement include Caldas (1993) who studied all public schools in Louisiana. Achievement was not related to school size except for central city schools where larger schools were linked to lower achievement. Franklin and Crone (1992) and Hoagland (1995) found school size to have little influence on student achievement except in conjunction with SES. Ramirez (1992) conducted a recent review of school size and achievement literature and concluded that little difference was present between the effects of large and small schools on student achievement. He stated that school size research may be asking the wrong question and that the relationship to a school's community should be the focus of study.

In contrast to these findings, researchers have found larger school size to be positively related to student achievement in additional studies. A Governor's Commission on Schools (Committee on School Organization, 1973) in Illinois reported that small size school districts were inadequate and could not provide the necessary range of services. Huang and Howley (1993) studied schools in Alaska, and achievement and size were positively related in that state; however, the authors noted that smaller Alaskan secondary schools were farther from population centers, implying poverty and remote locations, and that the Alaskan population was atypical. Forbes, Fortune, and Packard (1993) considered student grades in individual high school courses in North Carolina secondary schools. Effects of school size varied by subject in this study, with students in larger schools being more likely to have higher grades in biology and physics than students in smaller schools. Harnisch (1987) reported that school size had a statistically significant, though small, effect on student achievement.

Research in which school size was negatively related to achievement has been reported throughout the literature. One meta-analysis of studies (Greenwald, Hedges, & Laine, 1996) related student inputs to student achievement through the use of school size as one of its variables. Based on 60 primary studies, these authors concluded that student achievement was negatively related to school size. Smaller schools had higher achievement in the full analysis and also in a subsample of 26 studies conducted since 1970.

When researchers used a multi-state population, results were similar even when SES was not considered. Walberg and Walberg (1994) studied the National Assessment of Educational Progress (NAEP) test scores of students from 37 states. They concluded that states with schools and school districts that were larger than the average size of schools and schools districts in the study had lower than average NAEP scores. Lee and Smith (1994) studied the scores of 22,000 eighth graders from the 1988 NELS study and the 1990 followup. Smaller school sizes were positively correlated to higher achievement. Fowler and Walberg (1991) reported similar results in their study of 293 secondary schools in New Jersey.

Two particularly important studies performed in individual states were important to the organization and to the method of this study. Friedkin and Necochea (1988) studied all schools in California and reported a weak positive correlation (.149) between student achievement and school size at grade 12 but weak negative correlations (-.198 to -.033) for grades 3, 6, and 8. When socioeconomic factors were held constant, however, school size was then inversely related to student academic performance. These authors determined that, though large schools could possibly have no effect or even a positive effect on students from higher socioeconomic levels, they had particularly negative effects on disadvantaged students.

Friedkin and Necochea (1988) quantified the effect of school size on students of lower SES. Negative effects of large schools on low SES students were determined to be significantly greater than positive effects of large schools on higher SES students. The effects remained whether SES was measured by student reports of parent education or teacher reports of parent occupations, and the outcome was verified for students in grades 3, 6, and 12. Hoagland (1995) conducted a similar study in California with the same results. In addition, Hoagland identified a significantly greater negative effect on reading scores than on scores for other subjects for low SES students in larger schools. Research up to this point could provide support for the conclusion that larger schools were not harmful to achievement for higher SES students and might have positive influences on achievement. For low SES students, however, large schools appear to have a negative effect on academic achievement.

Howley's replication (1995) of the work of Friedkin and Necochea led to the same conclusions for schools and students in West Virginia. Though Howley noted that California and West Virginia are very different states, he found the same lack of statistical significance in the relationship between school size and student achievement for grades 6, 9, and 11. Students in grade three in larger schools had higher achievement test scores than students in other grade levels. When SES was considered, however, the results differed. In grades 6, 9, and 11, statistically significant negative relationships existed between school size and achievement. The works of Friedkin and Necochea and of Howley lend strong direction to the conduct of future school size studies.

Because socioeconomic status was a significant variable in many studies already discussed, those studies will be discussed in more detail. Social economic variables include socioeconomic status (SES) of the students enrolled in a school and SES of their parents and community. The effect of SES is particularly important in educational research because the socioeconomic status of students in a school is a factor that cannot be manipulated.

Fowler and Walberg (1991) made a case for considering school size as an equity issue based on the differences in effect of school size on students of differing socioeconomic levels. If the students most adversely affected by large school size are those students who are most at-risk, the SES literature must be carefully explored. Effects of school size on the achievement of low socioeconomic status students have been the focus of a number of studies. Several studies in that area (e.g., Friedkin & Necochea, 1988; Hoagland, 1995; Howley, 1995) were reviewed in relation to school size and achievement. Results from those studies that are salient to this section are reiterated below, and additional studies are presented.

In a study particularly designed to determine whether effective schools are equally effective for students without regard to SES, only schools which performed well on the 1988 National Educational Longitudinal Study (NELS) were included (Crone & Tashakkori, 1992). Using data from 989 schools that were considered effective, based on achievement test results, variance of student achievement was examined. Schools with lower SES had a significantly higher variance in student achievement than schools with higher SES. These results provide evidence for the importance of considering the variance of student achievement as it relates to SES and school effectiveness.

A similar study involved 1336 schools in Louisiana (Franklin & Crone, 1992). The researchers considered school size as one of the variables along with student achievement and SES. School size had little impact on test scores except in conjunction with SES. Franklin and Crone concluded that large schools were not effective for economically deprived students.

The work of Friedkin and Necochea (1988) and Hoagland (1995) in California, of Franklin and Crone (1992) in Louisiana, of Kearney (1994) in Idaho, and of Howley (1995) in West Virginia all reached essentially the same conclusions based on statistically significant relationships among school size, student academic achievement, and SES. In all studies, students of lower socioeconomic status had lower achievement in larger schools than students of lower socioeconomic status in smaller schools. In our estimation, Howley (1995) stated well the conclusion of these studies that "the direct association of size and achievement is neither practically nor statistically significant, but, instead socioeconomic status governs the relationship" (p.19).

A strong relationship among student socioeconomic status, student academic achievement, and school size appears to be present in the research literature. All current research supported the conclusion that larger schools have a negative effect on the academic achievement of low SES students. Any work with optimal school size must consider the socioeconomic status of each school's population.

An additional problem in evaluating and applying school size research results from the legal authority for education in the United States. Because education is constitutionally the province of each state, state-to-state differences exist in such major factors as funding, graduation requirements, curriculum, administration, local authority, and activities. All of these factors affect schools and may affect schools of different sizes in different ways. Though some generalizations may be possible from one state to another, conclusive research relating size and quality must be specific to the state where it is to be applied, to that state's funding, and to the division of power between that state and its local education agencies.

Clear findings are not present on the relationship between school size and achievement, and the current relationship of these factors to SES has not been studied specifically in the state of Georgia. Thus, the established arguments for larger schools and school consolidations need to be reconsidered. Our purpose in conducting this study was to determine the relationship between secondary school size and student achievement for schools in the state of Georgia.

Methods and Procedures

All Georgia public secondary schools were included in this study with the exception of night schools, alternative schools, magnet schools, and adult schools. In accordance with the methodology established in previous studies, we placed schools in size categories according to enrollment. Several researchers have used size categories including Monk (1987) who divided New York school districts into 10 groups of districts according to size. Monk and Haller (1993) designated small, medium, and large categories by the number of seniors in the graduating class, as did Barker and Gump (1964). Pethel's study of Georgia schools (1978) used five size categories. Hoagland (1995) categorized California schools into seven size categories and three SES strata.

Each public secondary school in Georgia was designated as a small (i.e., schools with an enrollment of 600 or fewer students), medium (i.e., schools with an enrollment from 601 to 1000 students), or large school (i.e., schools with an enrollment of 1001 or more students). Data for this categorization were obtained from the Georgia Public Education Report Card (Georgia Department of Education, 1996). The figures were provided to the Georgia Department of Education by each Georgia school system.

Our dependent variable was academic achievement. School percentile scores on Reading and Mathematics on the Test of Academic Proficiency (TAP), administered to a matrix sample of eleventh grade students at all Georgia public secondary schools and reported in the Georgia Public Education Report Card (Georgia Department of Education, 1996), composed achievement on a nationally normed test. The TAP scores were converted to *z* scores for data analysis. Percent of students passing the writing, English, and mathematics portions of the Georgia High School Graduation Test (GHSGT) provided another measure of achievement.

Table 1
Means, Standard Deviations, Minimum and Maximum Values, and Number of Subjects for Enrollment and % Free or Reduced Lunch in the 1994-1995 School Year

Item by School Size	*M*	*SD*	*n*	Min	Max
Enrollment					
Small	421.99	120.65	74	127	600
Medium	808.22	116.27	78	604	1000
Large	1425.75	317.02	157	1003	2450
Total	1029.48	488.71	309	127	2450
% Free or Reduced Lunch					
Small	47.91	19.82	74	6.95	88.96
Medium	35.03	17.87	78	4.11	84.33
Large	26.68	18.21	157	.52	84.11
Total	33.87	20.38	309	.52	88.96

Validity and reliability information on the GHSGT were reported by Bunch and Klaric (1997). Kuder-Richardson Formula 20 reliabilities on the English (0.80) and on the Mathematics (0.92) subtests were generally high. Measurement of test validity was intended to determine whether the GHSGT subtests measured the instruction actually provided in Georgia schools. Evidence present in the GHSGT manual was consistent with what Georgia students should know and be able to do; test forms were consistent and comparable from year to year; and the test was reported to be free of bias based on gender or race. Bunch and Klaric (1997) reported the procedures for identification of test objectives by educators and construction of test items by the test development contractor. To determine consistency over time, editions of the test were statistically equated to previous editions, affording students taking each edition an opportunity to pass equal to that of students in previous test administrations. Criterion and construct validity were established using statistically significant correlations between course grades and subtest scores (*r*s of .45 and .46 on English and mathematics, respectively).

Use of both nationally normed and state competency test scores is supported in the literature. Researchers have used standardized test scores only (e.g., Lee & Smith, 1994; Walberg & Walberg, 1994), state assessment program scores only (e.g., Friedkin & Necochea, 1988; Stecklenberg, 1991) or some combination of the two scores (e.g., Howley, 1995). Scores from the TAP and GHSGT were recorded from the state Report Card. Both tests were chosen because they were administered to the same group of students and under standardized conditions. Similar to Lee and Smith (1994), Walberg and Walberg (1994), Friedkin and Necochea (1988), and Stecklenberg (1991), we analyzed the TAP and GHSGT scores separately, rather than using a composite variable, to avoid confounding the results of a norm-referenced test (TAP) and the results of a criterion-referenced test (GHSGT).

To allow for some control of socioeconomic status as an intervening variable, the percent receiving free or reduced lunch at each Georgia secondary school was obtained from the Georgia Public Education Report Card. Socioeconomic status was recorded from the Georgia Public Education Report Cards as a total of the two reported percents of students eligible for free lunch and eligible for reduced price lunch. Though Howley (1995) raised concern about the use of this information as an indicator of SES, no other good alternative for consideration of data on a school level was available in his West Virginia study, and a similar situation exists in Georgia. Warnock (1987) used free and reduced price lunch percents as the SES indicator in his study of dropout rates and achievement in Georgia. Poverty data for counties would not allow for school level distinctions. Acknowledging that the total percent of a school's students who are eligible for free or reduced price lunch is a convenient proxy for SES and has recognized problems, that value will represent a school's SES for that particular school year.

Data Analysis

Following the computation of basic statistics including means, standard deviations, and minimum and maximum values, a multivariate analysis of variance (MANOVA) was conducted to determine whether statistically significant differences existed on the GHSGT and on the TAP among the groups considered in this study. This test was followed by univariate analysis of variance (ANOVAs) to compare test scores by school size category while controlling for percent receiving free or reduced price lunch. Significant *F* statistics were analyzed by multiple post hoc comparisons using Scheffé to identify the within-groups differences.

Results

Means, standard deviations, and minimum and maximum values for the variables in this study are reported in Tables 1, 2, and 3. Minimum and maximum values were included to provide the range of each of the variables. Small schools had an average of approximately 20% more students receiving free or reduced price lunches than large schools (47.91% compared to 26.68%), and the mean percent for medium schools was almost halfway between them at 35.03%.

On the TAP tests (see Table 2), all size categories performed more poorly on Reading than on Math. Test scores on the GHSGT subtests (see Table 3) also followed the same pattern for all size categories. All groups had higher mean scores on the English subtest followed by the Writing subtest, and all groups had their lowest mean percent passing on the Math test. Standard deviations of the test scores of the small schools category were larger than the standard deviations of the other groups and also larger than the standard deviation of the total group on every test score. Based on that measure of variability, the dispersion of scores from high to low was greater in the small schools category than in the other two groups.

Table 2
Means, Standard Deviations, Minimum and Maximum Values for TAP Test Scores in the 1994-1995 School Year

Test by School Size	*M*	*SD*	*n*	Min	Max
TAP Math *z* scores					
Small	-.07	.34	74	-.81	.55
Medium	.11	.27	76	-.61	1.04
Large	.18	.33	157	-.50	.95
Total	.10	.33	307	-.81	1.04
TAP Reading *z* scores					
Small	-.30	.38	74	-1.08	.77
Medium	-.18	.31	76	-.88	.55
Large	-.10	.34	157	-.84	.64
Total	-.17	.35	307	-1.08	.77

Note: TAP test scores for two medium schools were not available.

Table 3
Means, Standard Deviations, Minimum and Maximum Values for GHSGT Test Scores for First Time Test Takers in the 1994-1995 School Year

GHSGT Test	*M*	*SD*	*n*	Min	Max
GHSGT Writing					
Small	84.99	9.55	74	41	100
Medium	88.14	6.62	78	71	100
Large	89.71	7.91	157	50	100
Total	88.18	8.24	309	41	100
GHSGT English					
Small	86.03	9.73	74	50	100
Medium	90.10	4.63	78	75	98
Large	91.70	5.20	157	76	100
Total	86.94	6.84	309	50	100
GHSGT Math					
Small	76.82	12.49	74	42	100
Medium	82.76	7.65	78	64	98
Large	85.40	8.96	157	60	99
Total	82.67	10.21	309	42	100

BEST COPY AVAILABLE

To determine whether a statistically significant difference was present among the GHSGT and TAP achievement test scores as a function of school size, a MANOVA was conducted and was found to yield statistically significant results, $p < .001$. Accordingly, univariate analyses of variance were examined for each of the five sets of test scores with school size categorized as small, medium, or large. The mean scores among the school size categories were statistically significantly different on TAP Math, $F(2, 306) = 15.54$; TAP Reading, $F(2, 306) = 8.79$; GHSGT Writing, $F(2, 308) = 8.66$; GHSGT English, $F(2, 308) = 19.40$; and GHSGT Math, $F(2, 308) = 19.86$, $ps < .01$. Multiple post hoc comparisons using Scheffé were used to identify the within-groups differences.

On the TAP Math and Reading subscales, the small school mean scores were significantly lower than large school mean scores with mean differences of .25 and .20 respectively. Medium school mean scores were between the large and small schools on both measures but were not significantly lower than the large school mean scores on either measure. Medium school mean scores were significantly higher than the scores of small schools on the TAP Math but not on the TAP Reading.

The large school category had the highest mean scores on all three GHSGT tests, followed in order by the medium school category and then the small school group. On each test, small schools' scores were significantly lower than those of large schools, with mean differences for Writing, English, and Math of 4.72%, 5.67%, and 8.57% respectively. For both English and Math, medium schools and large schools formed a homogeneous subset, but in the area of writing, medium school scores were not significantly higher than small school scores.

Because previous researchers (Forbes, Fortune, & Packard, 1993; Friedkin & Necochea, 1988) had linked socioeconomic factors to achievement, a multivariate analysis of variance comparing test scores by school size category was conducted controlling for the percent of students who receive free or reduced price lunches. No statistically significant between-subjects effects were found for any tests: TAP Math, $F(2, 306) = .44$; TAP Reading, $F(2, 306) = .33$; GHSGT Writing, $F(2, 306) = .80$; GHSGT English, $F(2, 306) = 1.16$; and GHSGT Math, $F(2, 306) = .33$, $ps > .05$.

The relationship between school size and achievement was also examined using correlations. Bivariate correlations of interest were the correlations of school size and test scores, with the percent free/reduced lunch uncontrolled and controlled for in the statistical analyses. Relationships of the TAP and GHSGT test scores with school size were statistically significant, ranging from .23 to .33, when percent free/reduced lunch was not considered in the analysis. When partial correlations controlled for the percent of students receiving free or reduced price lunches (see Table 4), school size categories were not statistically significantly related to the test score variables.

Table 4
Correlations of School Size and Test Scores for Georgia Secondary Schools, Not Controlling and Controlling for % Free/Reduced Lunch

	School Size	
Test Scores	Not Controlling for % Free/Reduced Lunch	Controlling for % Free/Reduced Lunch
TAP		
Math	.30*	.02
Reading	.23*	-.04
GHSGT		
Writing	.23*	-.07
English	.33*	.07
Math	.33*	.02

* Correlations statistically significant at the .01 level

Discussion

Two tentative conclusions appear to be present in our findings. First, in considering student academic achievement, secondary school size in Georgia was directly related to achievement. Students in small schools exhibited poorer academic achievement than did students enrolled in larger schools. The poorer academic achievement was found on both a norm-referenced and on a criterion-referenced measure. These findings were consistent with Stecklenberg's (1991) study and Harnisch's (1987) research in which those authors reported a slightly positive correlation between academic achievement and school size. Results herein are also similar to the findings of Friedkin and Necochea (1988) and Hoagland (1995) who found weak positive correlations between school size and achievement, and who then went on to consider socioeconomic factors.

Second, when socioeconomic factors were considered in the analysis of the 1996 Georgia Public Education Report Card, students did not differ in their academic achievement as a function of school size. That is, when we controlled for the percent of students receiving free or reduced lunch in our statistical analysis, no statistically significant differences in student achievement were found. These findings again parallel the findings of previous researchers (e.g., Franklin & Crone, 1992; Friedkin & Necochea, 1988; Hoagland, 1995; Howley, 1995). Therefore, the issue when considering student academic achievement by school may not be school size but rather factors related to poverty. Unfortunately, this statement is made tentatively because of the problems in using free or reduced lunch as a convenient proxy for socioeconomic status (Howley, 1995).

Findings in our study were not consistent with the results of other researchers who found a negative relationship between school size and achievement, when socioeconomic status factors were not considered (e.g., Fowler & Walberg, 1991; Greenwald, Hedges, & Laine, 1996; Lee & Smith, 1994). Unfortunately, only a general comparison of findings across studies is possible, due to differences in research methods and procedures, populations, and specific tests used as indicators of academic achievement. Additionally, results of our research did not support the work of investigators who found no statistically significant relationship between school size and academic achievement (e.g., Caldas, 1993) or investigators who found mixed results for different subject areas (e.g., Forbes, Fortune, & Packard, 1993).

Lest readers overgeneralize our findings, several caveats are in order. First, our findings are based on students enrolled in Georgia public schools, schools which were placed into three categories for statistical analysis. Although this categorization was consistent with the way in which previous studies (e.g., Barker & Gump, 1964; Hoagland, 1995; Monk, 1987) have been conducted, other researchers might incorporate school size as a continuous variable into their statistical analyses. Second, academic achievement was defined by two standardized tests, one of which is a Georgia standardized criterion-referenced test. Whereas we decided to analyze each test separately, similar to previous researchers (e.g., Lee & Smith, 1994; Stecklenberg, 1991; Walberg & Walberg, 1994), an analysis of a composite variable of the two measures might be of interest to researchers (e.g., Howley, 1995). Third, socioeconomic status was defined solely by free or reduced lunch enrollment which is only one of several ways by which socioeconomic status can be defined. Although a convenient proxy for SES, readers should recognize the substantial problems with its use. Fourth, the data we analyzed were obtained from the 1994-1995 Georgia Public Education Report Card, and the extent to which these findings would be replicable across other years of data is unknown. Therefore, we urge readers to be cautious in any generalizations they might make based upon our findings.

Clearly, more research is needed on school size and student achievement, not only on standardized tests, but on other measures and for other constructs as well. That is, student outcomes such as civic involvement and citizenship are important ones that need to be investigated as a function of school size. Because prior researchers have reported differing effects of school size on students from differing socioeconomic levels, the factor of socioeconomic status defined in a more objective and valid way than by free or reduced lunch should be considered. Researchers should investigate school size and student achievement in states with student populations that differ from student populations in Georgia. Moreover, consideration of longitudinal achievement data on achievement may provide a more accurate measure of the relationship of the variables in this study. Should future research findings replicate findings reported herein, policy makers would be in stronger positions to make decisions about school size and school consolidation issues. Until such time, policy makers would be advised to view the school size research critically and recognize the myriad of difficulties in arriving at fixed conclusions.

References

Barker, R. G., & Gump, P. V. (1964). *Big school, small school.* Palo Alto, CA: Stanford University Press.

Bunch, M. B., & Klaric, J. S. (1997). *Georgia high school graduation tests: Reliability and validity.* Atlanta: Measurement Incorporated.

Burrup, P. E., Brimley, V., & Garfield, R. R. (1988). *Financing education in a climate of change.* 4th ed. Boston: Allyn and Bacon.

Caldas, S. J. (1993). Reexamination of input and process factor effects on public school achievement. *Journal of Educational Research, 86,* 206-214.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Department of Health, Education, and Welfare.

Committee on School Organization. (1973). *Opportunities for excellence: Findings, conclusions, and recommendations of a survey of Illinois school district organization. Final report of the committee on school organization.* Springfield, IL: Governors Commission on Schools. (ERIC Document Reproduction Service No. ED 091 831)

Conant, J. B. (1959). *The American High School Today.* New York: McGraw-Hill.

Costenbader, V. K., & Markson, S. (1994). School suspension: A survey of current policies and practices. *NASSP Bulletin, 78,* 103-107.

Cotton, K. (1997). School size, school climate, and student performance. [On-line]. *School Improvement Research Series,* Close-up #20. Available: http://www.nwrel.org/scpd/sirs/10/c020.html.

Crone, L. J., & Tashakkori, A. (1992, April). *Variance of student achievement in effective and ineffective schools: Inconsistencies across SES categories.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 346 613).

Forbes, R. H., Fortune, J. C., & Packard, A. L. (1993, February). *North Carolina rural initiative study of secondary schools: Funding effects on depth of the curriculum.* Paper presented at the annual meeting of the Eastern Educational Research Association.

Fox, W. F. (1981). Reviewing economies of size in education. *Journal of Education Finance*, 6, 273-296.

Fowler, W. J., Jr., & Walberg, H. J. (1991). School size, characteristics, and outcomes. *Educational Evaluation and Policy Analysis, 13*, 189-202.

Franklin, B. J., & Crone, L. J. (1992, November). *School accountability: Predictors and indicators of Louisiana school effectiveness.* Paper presented at the annual meeting of the Mid-South Educational Association, Knoxville, TN. (ERIC Document Reproduction Service No. ED 354 261)

Friedkin, N. E., & Necochea, J. (1988). School system size and performance: A contingency perspective. *Educational Evaluation and Policy Analysis, 10*, 237-249.

Georgia Department of Education. (1996). *1994-1995 Georgia public education report cards: State, systems and schools* [CD-ROM]. Atlanta, GA: Author.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*, 361-396.

Haller, E., Monk, D., & Tien, L. (1993). Small schools and higher-order thinking skills. *Journal of Research in Rural Education, 9*, 66-73.

Harnisch, D. L. (1987). Characteristics associated with effective public high schools. *Journal of Educational Research, 80*, 233-240.

Hoagland, J. P. (1995). *The effect of high school size on student achievement as measured by the California Assessment Program.* Unpublished doctoral dissertation, University of La Verne, CA.

Howley, C. (1995, November 15). The Matthew principle: A West Virginia replication. [27 p.] *Educational Policy Analysis Archives* [On-line serial], *3*(18). Electronic journal:
http://olam.ed.asu/epaa/v3n18.html

Huang, G., & Howley, C. (1993). Mitigating disadvantage: Effects of small-scale schooling on student achievement in Alaska. *Journal of Research in Rural Education, 9*, 137-149.

Lee, V. E., & Smith, J. B. (1994). Effects of high school restructuring and size on gains in achievement and engagement for early secondary school students. (Grant No. R117Q00005-94). Washington, D. C.: Center on Organization and Restructuring of Schools, supported by the U. S. Department of Education, Office of Educational Research and Improvement.

McGuffey, C. W. (1991). Large school/small school issue: Georgia schools. *CEFPI Educational Facility Planner, 29*, 17-24.

McGuffey, C. W., & Brown, C. L. (1978). The relationship of school size and rate of school plant utilization to cost variations of maintenance and operation. *American Educational Research Journal, 15*, 373-378.

Monk, D. H. (1987). Secondary school size and curriculum comprehensiveness. *Economics of Education Review, 6*, 137-150.

Monk, D. H., & Haller, E. J. (1993). Predictors of high school academic course offerings: The role of school size. *American Educational Research Journal, 30*, 3-21.

Pethel, G. E. (1978). *An investigation of the relationship of school size and program quality in the public high schools of Georgia.* Unpublished doctoral dissertation, University of Georgia, Athens.

Pittman, R. B., & Haughwout, P. (1987). Influence of high school size on dropout rate. *Educational Evaluation and Policy Analysis, 9*, 337-343.

Ramirez, A. (1992). Size, cost, and quality of schools and school districts: A question of context (Report No. RC 019 322). In *Source Book on School and District Size, Cost, and Quality* (Report No. RC 019 318). Minneapolis, MN: Minnesota University, Hubert H. Humphrey Institute of Public Affairs; Oak Brook, IL: North Central Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED 361 162)

Sher, J. P., & Tompkins, R. B. (1977). Economy, efficiency, and equality: The myths of rural school district consolidation. *CEFP Journal, 15*, 4-14.

Stecklenberg, C. R. (1991). *The effects of public high school size on student achievement: A meta-analysis.* Unpublished doctoral dissertation, University of Georgia, Athens.

Swanson, A. D. (1988). The matter of size: A review of the research on relationships between school and district size, pupil achievement, and cost. *Research in Rural Education, 5*, 1-8.

Walberg, H. J. (1992). On local control: Is bigger better? (Report No. RC 019 324). In *Source Book on School and District Size, Cost, and Quality* (Report No. RC 019 318). Minneapolis, MN: Minnesota University, Hubert H. Humphrey Institute of Public Affairs; Oak Brook, IL: North Central Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED 361 164)

Walberg, H. J., & Walberg, H. J., III. (1994). Losing local control. *Educational Researcher, 23*, 19-26.

Warnock, C. M., Sr. (1987). *Student achievement and dropout rates in Georgia school districts.* Unpublished doctoral dissertation, University of Georgia, Athens.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 27-32

Fourth and Fifth Grade Students' Attitudes Toward Science: Science Motivation and Science Importance as a Function of Grade Level, Gender, and Race

John R. Slate
Valdosta State University

Craig H. Jones
Arkansas State University

Because some scientists have expressed a concern that poor science instruction in elementary school is causing students to develop negative attitudes toward science, we surveyed 941 fourth and fifth graders in a school district in the Southeastern United States. These students expressed positive attitudes toward science on 20 of 21 questionnaire items. Principal factor analysis with Varimax rotation revealed two underlying dimensions: perceived importance of science and motivation to study science. Students perceived science to be moderately important and were moderately motivated to study it. No gender differences were found, but fifth grade students indicated slightly less positive attitudes than did fourth grade students. In addition, African-American students had slightly less favorable attitudes than did White students. Implications are discussed.

Many scientists have become concerned about growing negative public attitudes toward science in our society. Theocharis and Psimopoulos (1987) spoke for many scientists when they lamented a perceived devaluation of science in our society. Indeed, scientists have become so concerned about antiscience attitudes that the New York Academy of Sciences held a special conference on this topic entitled "The Flight From Science and Reason" (Rios, 1995).

Three basic factors are typically perceived to be the causes of antiscience attitudes. One of these factors is the rise of social relativism and postmodernism in academic circles (Gross & Levitt, 1994; Lederman, 1996; Schick, 1997). For example, Feyerabend (1975) argued that science is a religious ideology, and therefore, does not provide an objective method for determining truth. Other postmodern writers have blamed science for most, if not all, of the problems of contemporary society (Englebretsen, 1995).

A second perceived cause of antiscience attitudes is the fact that the mass media has increasingly become a purveyor of pseudoscience (Lederman, 1996) and negative images of scientists (Evans, 1996). Television networks run numerous programs that support pseudoscientific beliefs such as "The X-Files," "Alien Autopsy," and "The Mysterious Origins of Man." Furthermore, Gerbner (1987) found that scientists are more likely to be killed in television programs than are the members of any other profession, and scientists are also more likely to engage in actions that result in the deaths of others. Gerbner also found that the more television people watched, the more likely they were to believe that science was a dangerous and undesirable occupation. In addition, Evans (1996) noted that the mass media increasingly portrays science as a useless approach to understanding reality. In many movies and television programs characters who take rational, scientific views of the world are portrayed as foolish and unable to resolve problems, whereas believers in pseudoscience and the supernatural are portrayed as wise and as the heroes who ultimately save the day. Finally, Wiseman and Jeffreys (1997) found that 85.2% of the passages they examined in five "nonfiction" children's books on paranormal phenomena endorsed pseudoscientific beliefs.

The third perceived cause of antiscientific attitudes is the educational system. Schools are blamed, in part, for failing to provide sufficient scientific literacy to combat the negative and inaccurate views of science portrayed by the media. Lederman (1996), for example, argued that the state of science education is so bad that it is placing our society at risk. Elementary school teachers are often singled out for particular criticism because, unlike secondary school science teachers, elementary teachers do not major in science. For example, Padian (1993) stated that because "a majority of elementary-school teachers

---

John R. Slate, Ph.D. is Professor in the Department of Educational Leadership at Valdosta State University. He is currently involved in a P-16 initiative and in the Southeastern Rural Systemic Initiative funded by the National Science Foundation. Craig H. Jones, Ed.D. is Professor in the Department of Counselor Education and Psychology at Arkansas State University. Correspondence regarding this article should be directed to John R. Slate, Valdosta State University, Department of Educational Leadership, Valdosta, GA 31698-0090 or via e-mail to jslate@valdosta.edu.

640

have poor to nonexistent backgrounds in science . . . it is optimistic to think that they can transmit science effectively to students" (p. 388). Kepler (1996) described a lack of appropriate science background, and a resulting lack of comfort in teaching science, as an important obstacle to science teaching in elementary school. Kepler also noted that some elementary school teachers do not even like science, quoting one teacher as saying "I didn't like science when I was in school . . . . The main thing I remember was being forced to cut up a frog" (p. 46). Thus, elementary school teachers may directly convey a dislike of science to students.

Previous research has documented that students with negative attitudes toward science are less likely to take science courses (Gabel, 1981), and they demonstrate lower achievement in the courses they do take (Oliver & Simpson, 1988), than do students with positive attitudes toward science. Thus, for science education in the United States to improve significantly, students must hold positive attitudes toward science and believe that engaging in science is a valuable and rewarding activity. If students hold the negative attitudes toward science attributed to them by many scientists, students are unlikely to have the motivation to study science seriously or to aspire to careers in science. The argument that antiscience attitudes are widespread, however, is mainly anecdotal, and some surveys have indicated that adults, although often ignorant of scientific concepts, do not have negative attitudes (Frazier, 1996). Thus researchers need to ascertain the attitudes students have toward science, especially among elementary school students who purportedly are the most likely to receive poor science teaching from teachers who dislike science. Although declining attitudes during elementary school have been reported in the literature, these reports have been based primarily on retrospective accounts provided by secondary school and college students (Gabel, 1981; Gogolin & Swartz, 1992; Oliver & Simpson, 1988; Yager & Penick, 1984). Weinburgh (1994), however, did compare the attitudes toward science of fourth grade students with the attitudes of seventh and tenth grade students. Weinburgh found that students' attitudes toward science became increasingly negative across grade levels and that, regardless of grade level, boys had more positive attitudes toward science than did girls.

The present study was conducted to ascertain the attitudes of fourth and fifth grade students toward science. The specific research questions addressed were: (a) Do fourth and fifth grade students have negative attitudes toward science? and (b) Do students' attitudes toward science differ as a function of their gender, race, and grade level?

Method

Participants

Participants were 941 elementary school students in a community of approximately 45,000 residents located in the southeastern United States. Surveys were conducted as part of a project implemented by the school to improve science instruction in the upper elementary grades. Students in all fourth and fifth grade classes in the school district were surveyed. Thus the sample included every student at these grade levels who was in school on the day of the survey.

The total sample included 423 fourth graders, 516 fifth graders, and 2 students who did not report their grade level. There were 444 males, 494 females, and 3 students who did not report their gender. The sample was predominantly African American (n = 646) with a substantial number of Whites (n = 230), and a few Asian (n = 13) and Hispanic (n = 8) students. The remaining 44 students either listed their ethnic background as other (n = 21) or did not respond to this item (n = 23).

Procedure

All students completed a 24-item questionnaire, titled "Attitudes Toward Science Survey," which was adapted specifically for this study from surveys used by Gogolin and Swartz (1992) and Weinburgh (1994). Gogolin and Swartz (1992) developed their attitude inventory to assess attitudes toward science of post-secondary students, and Weinburgh (1994) revised this instrument to measure attitudes toward science in fourth, seventh, and tenth grade students. There were 15 items worded so that agreement indicated a positive attitude toward science and 9 items reverse worded so that disagreement indicated a positive attitude toward science. The scale was further modified in the present study by deleting three items from the analyses as described below.

The questionnaires were administered by the students' regular classroom teachers. The teacher wrote the instructions for completing the questionnaire on the board, and distributed a written copy of the questionnaire and a Scantron sheet to each student. The teacher then read aloud the instructions and all of the questionnaire items. Students responded to each item immediately after it was read by the teacher. The teacher waited until all students had responded before reading the next item. These responses were made by darkening spaces on the Scantron sheet using a four point scale from *strongly disagree* to *strongly agree*.

Items on which disagreement indicated a positive attitude toward science were reverse scored, and all responses were summed to produce a single overall attitude toward science scale. The higher a student's score on this scale, the more positive was his or her overall attitude toward science. The appropriateness of summing all responses into a single scale was assessed using Cronbach's coefficient alpha. As a result of this analysis, three items were deleted from the analysis because they had negative item-total correlations and, therefore, did not assess attitudes toward science in the way intended by the authors of the scale. The coefficient alpha of the resulting 21 item scale was .84 indicating very high internal consistency. This scale was labeled the General Attitude scale, and the 21 items on this scale are presented in Table 1.

Table 1
Percentage of Total Sample Indicating a Positive Attitude Toward Science on Each Questionnaire Item

Questionnaire Item	Percent Positive
1. Science is useful for the problems of everyday life	64.7
2. Science is something which I enjoy very much	70.1
3. I like the easy science assignments best.	74.2
4. Doing science labs is fun.	90.8
5. I would like to do some outside reading in science.	69.3
6. There is little need for science in most jobs.	49.3
7. Most people should study some science.	79.4
9. Sometimes I read ahead in our science book.	60.4
10. Science is helpful in understanding today's world	82.4
11. I do not like anything about science	80.1
12. Science is of great importance to a country's development.	76.8
13. It is important to know science in order to get a good job. 71.2	
15. I enjoy talking to other people about science.	60.4
16. I would enjoy watching a science program on television.	76.8
17. I like the challenge of science assignments.	63.2
18. You can get along perfectly well in everyday life without science.	70.5
19. I would rather be told scientific facts than find them out from experiments.	70.8
20. Most of the ideas in science are not very useful.	72.4
22. It is important to me to understand the work I do in the science class.	84.6
23. Science is one of my favorite subjects.	64.3
24. I have a real desire to learn science.	72.2

Note. Items 8, 14, and 21 were excluded from the analysis.

Although the General Attitude scale score had high internal consistency, corrected item-total correlations ranged from .11 to .63 which suggested that students' general attitude consisted of a number of separate attitudes toward more specific aspects of science. Thus, a principal factor analysis with Varimax rotation was conducted to identify any component attitude dimensions. This analysis revealed three factors that had an eigenvalue greater than 1.00 and that accounted for a combined 27.74% of the variance.

The first factor had an eigenvalue of 2.37 and accounted for 11.3% of the variance. There were five items (i.e., 2, 15, 17, 23, and 24) with loadings greater than .4 on this factor. Because these items reflected students' motivation to engage in science, the first factor was labeled the Science Motivation scale. The coefficient alpha for this scale was .81, indicating high internal consistency.

The second factor had an eigenvalue of 2.10 and accounted for 10.0% of the variance. There were five items (i.e., 1, 7, 10, 12, and 13) with loadings greater than .4 on this factor. Because these items reflected how important students' believed science to be, the second factor was labeled the Science Importance scale. The coefficient alpha for this scale was .63, indicating sufficiently high reliability for research with grouped data (Salvia & Ysseldyke, 1991).

The third factor, with an eigenvalue of 1.34, accounted for 6.4% of the variance and was comprised of three items (i.e., 11, 19, and 20). Because of the low internal consistency of this factor (i.e., .48), this factor was dropped from further analyses.

Results

The first research question asked whether or not fourth and fifth grade students have negative attitudes toward science. Table 1 displays the percentage of students indicating a positive attitude toward science on each questionnaire item. Students indicated positive attitudes on 20 of the 21 items. The mean score on the General Attitude scale was 62.1 (*SD* = 10.7, range = 27 to 84) which is significantly above, $t(940) = 27.36, p < .001$, the scale midpoint of 52.5. In addition, 757 (80.5%) students had scores above 52.5, whereas only 183 (19.5%) had scores at or below 52.5. Thus, the students as a group had moderately positive overall attitudes toward science. Students had a mean score of 14.3 (*SD* = 4.1) on the Science Motivation scale which was above the scale midpoint of 10 (range = 5 to 20). Thus, students were moderately motivated to pursue science activities. The mean score on the Science Importance scale was 15.2 (*SD* = 3.1) which was above the scale midpoint of 10 (range = 5 to 20). Thus, students believed that science was moderately important.

The second research question asked whether or not differences in science attitudes existed as a function of students' gender, race or grade level. Because a sufficient number of Asian and Hispanic students were not present to include them in any analyses involving race, grade level and race effects were examined separately so that Asian and Hispanic students could be included in the grade level analysis. First, a 2 X 2 analysis of variance (ANOVA) of

BEST COPY AVAILABLE

General Attitude scores was conducted to determine whether differences were present in overall attitudes toward science as a function of gender and grade level. A statistically significant main effect was found, $F(1, 854) = 19.62, p < .0001$, for grade level with fourth graders ($M = 64.0$) having more positive attitudes than did fifth graders ($M = 60.7$). The main effect for gender, $F(1, 854) = 0.01, p > .05$, and the gender by grade level interaction, $F(1, 854) = 2.77, p > .05$, were not statistically significant. Two additional ANOVAs revealed that the difference in General Attitudes found for fourth and fifth grade students were also reflected in their Science Motivation scores, $F(1, 912) = 18.2, p < .0001$, and in their Science Importance scores, $F(1, 915) = 13.2, p < .0001$. Fourth grade students reported more positive attitudes for both Science Motivation ($M = 14.9$) and Science Importance ($M = 15.6$) than did fifth grade students (Ms = 13.8 and 14.9 respectively). There was no main effect for gender in either Science Motivation or Science Importance scores, and neither gender by grade interaction was statistically significant.

Because statistically significant differences in attitudes were found between the fourth and fifth grade students, a stepwise discriminant analysis of General Attitude scores was conducted with grade level as the criterion variable and the 21 survey items as the discriminating variables. The resulting discriminant function was statistically significant, $c^2(4) = 49.3, p < .0001$, and accounted for 5.6% of the between-groups variance (i.e., canonical correlation = .237). Six items contributed significantly to this function. These items and their standardized discriminant function coefficients are listed in Table 2. Group centroids were .27 and -.22 for the fourth and fifth grades respectively. Thus the positive discriminant coefficients indicate that fourth graders had more positive attitudes than did fifth graders on all the discriminating items. Four of these six items are from the Science Motivation subscale and indicate that fourth graders both enjoyed science activities more than did fifth graders and had a stronger desire to learn science. Fourth graders also saw science as more strongly related to getting a good job than did fifth graders.

The second research question was also addressed with a gender by race ANOVA on General Attitude scores. Race was treated as a dichotomous variable, African-American versus White, with Asian and Hispanic students excluded from the analysis due to their small numbers. Neither the main effect for gender, $F(1, 790) = 0.83$, nor the gender by race interaction, $F(1, 790) = 1.08$, was statistically significant. The main effect for race, however, was statistically significant, $F(1, 790) = 12.08, p < .001$, with White students having more positive attitudes than did African-American students (Ms = 64.2 and 61.3 respectively). Because the General Attitudes analysis had produced a statistically significant effect for race, gender by race ANOVAS were also conducted for the Science Motivation and Science Importance subscales. Effects were not statistically significant for either the Science Motivation scale or the Science Importance scale. Specific attitudinal differences as a function of race were identified with a stepwise discriminant analysis of the General Attitude items. The procedures for this discriminant analysis were the same as the procedures for the grade level analysis. The resulting discriminant function was statistically significant, $\chi^2(7) = 85.7, p < .0001$, and accounted for 10.3% of the between-groups variance (i.e., canonical correlation = .321). The seven items that contributed significantly to this equation are listed in Table 3 with their discriminant function coefficients. Because the group centroids were .55 for White students and -.21 for African-American students, positive coefficients indicate that White students had more positive attitudes than did African-American students, and negative coefficients indicate that African-American students had more positive attitudes than did White students.

Table 2
Items Discriminating Fourth and Fifth Grade Students

Item		Discriminant Coefficient
13.	It is important to know science in order to get a good job.	.62
15.	I enjoy talking to other people about science.	.43
16.	I would enjoy watching a science program on television.	.66
17.	I like the challenge of science assignments.	.69
23.	Science is one of my favorite subjects.	.42
24.	I have a real desire to learn science.	.46

Note. The positive coefficients indicate that fourth grade students expressed more positive attitudes than did fifth grade students.

Table 3
Items Discriminating African-American and White Students

Item		Discriminant Coefficient
9.	Sometimes I read ahead in our science book.	-.23
10.	Science is helpful in understanding today's world.	.33
13.	It is important to know science in order to get a good job.	-.33
15.	I enjoy talking to other people about science.	-.36
19.	I would rather be told scientific facts than find them out from experiments.	.74
20.	Most of the ideas in science are not very useful.	.27
23.	Science is one of my favorite subjects.	.30

Note. The positive coefficients indicate that White students expressed more positive attitudes than did African-Americans; negative coefficients indicate that African-American students expressed more positive attitudes than did Whites.

White students were more likely than were African-American students to include science as one of their favorite subjects. White students were also more likely to believe that science ideas are useful and helpful in understanding the world. The largest discriminating variable, however, was that White students were more likely to want to discover science facts for themselves, whereas African-American students were more likely to want to learn science facts by simply being told about these facts. On the other hand, African-American students were more likely than were White students to read ahead in their science books and to enjoy talking about science. African-American students were also more likely to believe that a knowledge of science is important to obtaining a good job.

Discussion

Although anecdotal evidence has raised considerable concern among many scientists that antiscience attitudes are rampant among the general public, surveys have not supported this concern. Just as a survey of adults did not find widespread negative attitudes toward science (Frazier, 1996), the present study found no evidence of widespread antiscience attitudes among upper elementary grade students. The absence of negative attitudes among these students is important because the elementary grades have been considered a breeding ground for antiscience attitudes by many writers (Kepler, 1996; Lederman, 1996; Padian, 1993), although the evidence for declining attitudes toward science during elementary school comes primarily from the retrospective reports of high school and college students (Gabel, 1981; Gogolin & Swartz, 1992; Oliver & Simpson, 1988; Yager & Penick, 1984). The only hint of a potential problem was a slight decline in attitudes among fifth graders. Research covering a wider range of grades will be needed to determine whether or not this decline is part of a more general trend. Indeed, longitudinal data in which students are followed across grade levels will be needed to resolve this issue. The present data, however, suggest no need to save students from the detrimental effects of elementary school science teachers.

Another interesting finding was that boys and girls did not differ in their attitudes toward science. Concerns have been raised because females are less likely to pursue advanced courses in mathematics and science than are males. Although females typically do not like math and science any less than do males, females typically think that learning math and science is less relevant to their futures (Kavrell & Petersen, 1984). In the present study, girls not only expressed the same general attitudes toward science as did boys but were equally motivated to study science. Girls also did not differ from boys in the perceived importance of science. Just as achievement differences in science between boys and girls seem to be disappearing (Slate, Jones, Turnbough, & Bauschlicher, 1994; Slate, Jones, Sloas, & Blake, 1997), attitude differences may be disappearing as well.

Students did express different attitudes toward science as a function of race. Although African-American students did not hold negative attitudes toward science, their attitudes were less positive than were the attitudes of White students. The magnitude of the difference, though statistically significant, was small but does raise important concerns because African-Americans demonstrate lower science achievement than other racial groups and are less likely to enter science-related careers (National Science Foundation, 1994).

Examination of the specific items that discriminated the attitudes of White and African-American students revealed a difference in the extent to which students had an intrinsic as opposed to extrinsic motivation to learn subjects. That is, more White students said that science was a favorite subject, that it was helpful in understanding the world in general, and that they enjoyed learning about science through experimentation. African-American students, on the other hand, saw science more as necessary to get a job and simply wanted teachers to tell them science facts. Because students who try to work out science problems show better transfer of knowledge than do students who simply try to remember information (Mayer & Wittrock, 1996), White students' typical attitudes are more compatible with achievement in science than are the typical attitudes of African-American students. In addition, scientists consider science to be a method of understanding the world rather than a mere body of factual knowledge (McCain & Segal, 1988). Thus, African-American students' desire simply to learn science facts could result in their failing to learn the most crucial aspects of science, that is, science methods. Therefore, elementary school science teachers need to make special efforts to involve African-American students in science activities in ways that stimulate their curiosity and demonstrate the value of using scientific methods to understand the world.

Although the present results are positive with regard to upper elementary grade students' attitudes toward science, and suggest that the effects of elementary school science education are not damaging as some scientists have argued, these findings must be interpreted with care. The students surveyed were all attending the same school district, therefore the attitudes expressed by these students may not be representative of other elementary grade students. In addition, the decline in attitudes displayed by fifth grade students might represent the beginning of a downward trend that could result in negative attitudes toward science developing in later grades. Clearly, additional research in other schools across a wider range of grade levels is needed to determine the generality of the current findings. Finally, the obtained differences between fourth and fifth grade students, and between African-American and White students, are correlational; and inferences regarding why such differences exist should be made with care.

References

Englebretsen, G. (1995). Postmodernism and New Age unreason. *Skeptical Inquirer*, *19*(3), 52-53.

Evans, W. (1996). Science and reason in film and television. *Skeptical Inquirer*, *20*(1), 45-48, 58.

Feyerabend, P. (1975). *Against method*. London: Verso.

Frazier, K. (1996). Where are the antiscience attitudes? Not among the general public. *Skeptical Inquirer*, *20*(6), 6-8.

Gabel, D. (1981). Attitudes toward science and science teaching of undergraduates according to major and the number of science courses taken and the effect of two courses. *School Science and Mathematics*, *1*, 70-76.

Gerbner, G. (1987, Spring). Science on television: How it affects public conceptions. *Issues in Science and Technology*, *3*, 109-115.

Gogolin, L., & Swartz, F. (1992). A quantitative and qualitative inquiry into the attitudes toward science of nonscience college students. *Journal of Research in Science Teaching*, *29*, 487-504.

Gross, P. R., & Levitt, N. (1994). *Higher superstition: The academic left and its quarrels with science*. Baltimore: Johns Hopkins University Press.

Kavrell, A., & Petersen, A. (1984). Patterns of achievement in early adolescence. *Advances in Motivation and Achievement*, *2*, 1-35.

Kepler, L. (1996). How to make hands-on-science work for you. *Instructor*, *105*(6), 46-52.

Lederman, L. M. (1996). A strategy for saving science. *Skeptical Inquirer*, *20*(6), 23-28.

Mayer, R. E., & Wittrock, M. C. (1996). Problem solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology*. (pp. 47-62). New York: Macmillan.

McCain, G., & Segal, E. M. (1988). *The game of science* (5th ed.). Pacific Grove, CA: Brooks/Cole.

National Science Foundation. (1994). *Women and minorities in science and engineering*. Washington, DC: Author.

Oliver, J. S., & Simpson, R. D. (1988). Influences of attitudes toward science, achievement motivation, and science self-concept on achievement in science: A longitudinal study. *Science Education*, *72*(2), 143-155.

Padian, K. (1993). Improving science teaching: The textbook problem. *Skeptical Inquirer*, *17*(4), 388-393.

Rios, E. (1995). "The Flight from Science and Reason": New York Academy of Sciences airs issues. *Skeptical Inquirer*, *19*(6), 42-44.

Salvia, J., & Ysseldyke, J. E. (1978). *Assessment in special and remedial education* (5th ed.). Boston: Houghton Mifflin.

Schick, T. (1997). The end of science? *Skeptical Inquirer*, *21*(2), 36-39.

Slate, J. R., Jones, C. H., Sloas, S., & Blake, P.C. (1997/98). Scores on the Stanford Achievement Test-8 as a function of sex: Where have the sex differences gone? *The High School Journal*, *81*(2), 82-86.

Slate, J. R., Jones, C. H., Turnbough, R., & Bauschlicher, L. (1994). Gender differences in achievement scores on the Metropolitan Achievement Test-6 and the Stanford Achievement Test-8. *Research in the Schools*, *1*, 59-62.

Theocharis, T., & Psimopoulos, M. (1987, October). Where has science gone wrong? [Commentary]. *Nature*, *329*, 595.

Weinburgh, M. (1994, March). *Achievement, grade level, and gender as predictors of student attitudes toward science*. Poster session presented at the annual meeting of the National Association of Research in Science Teaching, Anaheim, CA.

Wiseman, R., & Jeffreys, C. (1997). Bias and error in children's nonfiction books on the paranormal. *Skeptical Inquirer*, *21*(1), 41-43.

Yager, R. E., & Penick, J. E. (1984). What students say about science teaching and science teachers. *Science Education*, *72*(2), 143-155.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 33-42

Quantitative Graphical Display Use in a Southern U.S. School System

John V. Dempsey, Samuel H. Fisher, III, and Judith B. Hale
University of South Alabama

This paper reports the results of a survey of 429 teachers in an urban, racially mixed Southeastern school district. The survey elicited teacher perceptions of the value of using graphs and charts, when and how they taught and used graphical information, and how they themselves were trained in the use of graphical displays for instruction. Overall, the use of graphs is most prevalent in elementary schools and decreases as grade level increases. Although teachers perceive that students pay more attention to graphical information, most subject areas (excluding mathematics) report relatively infrequent use (one-third or lower) of these visuals in instruction. Teachers also perceived that it was more important to understand or use charts than to be able to construct them. Approximately one-half of the teachers surveyed reported receiving no training at all in the use of graphical displays. Findings are discussed with respect to Paivio's (1986a) dual coding theory.

In primary, middle, and secondary schools in the United States a vast amount of quantitative information is presented to students. Some of this information is presented in a body of work or in tables where the relationship between the numbers and the ideas is not clearly obvious. In other situations, graphical displays (charts, graphs, and related spatial or metaphoric representations of numeric data) are used in an effort to make quantitative relationships more concrete or easily interpreted.

Often, graphs are analyzed either by recognition (bottom-up processing) or by searching (top-down processing). Some tasks require a combination of recognition and search strategies (Brasell, 1990). Graphs may also be constructed by learners, either conventionally or using a graphing calculator or computer to construct graphs in "real time." These three actions that a learner may choose (i.e., recognize, search, or construct) are paralleled to some extent by the kinds of questions that a graph may be used to answer. Bertin (1973) and Wainer (1992) suggest that there are three levels of questions that a graph may answer: elementary level questions involving simple data extraction; intermediate level questions involving trends in the data; and overall level questions involving an understanding of the deep structure of the data.

Why Are Graphical Displays Used?

Graphical displays are used in many situations to represent large amounts of information concisely. They are particularly effective in showing intercomponent relationships and sequences (Moore, 1993). When used with more abstract textual information, they present a visual mode of information and therefore have the potential for encouraging dual coding (Paivio, 1983) or conjoint retention (Kulhavy, Lee, & Caterino, 1985). Dual coding theory, for example, suggests that information is represented in two fundamentally distinct systems. One of these systems is suited to verbal information and the other toward images. Paivio (1986a) suggests that incoming information can be coded in one or both systems. Information encoded in both systems would be enhanced compared to information encoded in only one of the systems. In addition, Paivio hypothesized that the nonverbal components of memory traces, which would include graphical displays, are often much stronger than verbal memories (Paivio, 1986b; Paivio & Csapo, 1975).

Some theorists suggest that graphical displays decrease processing demands in working memory that leaves cognitive resources for higher level operations such as the development of semantic macrostructures (Winn, 1991). Others promote the notion that understanding or constructing diagrams or graphs using learning strategies such as visual imagery provides a perceptual supplement for gaining insight into acquiring symbolic thinking essential for learning abstract concepts or mathematical problem-solving (Lin, 1979).

Certainly the use of graphical displays is widespread both in schools and in later life. Cleveland (1984) reports that approximately one-third of the space in some scientific journals is devoted to graphs. He emphasizes their importance by contending that readers who do not

---

John V. Dempsey is Associate Professor of Behavioral Studies and Educational Technology in the College of Education at the University of South Alabama. Samuel H. Fisher, III is Associate Professor of Political Science at the University of South Alabama. Judith B. Hale is a doctoral student in the Instructional Design and Development program at the University of South Alabama. Please address correspondence regarding the paper to John V. Dempsey, 3700 University Commons, University of South Alabama, Mobile, AL 36688 (e-mail: jdempsey@usamail.usouthal.edu).

scan scientific papers in detail are drawn toward graphs to extract information.

Whether graphical displays alone increase instructional effectiveness is debatable. Some researchers (e.g., Feliciano, Powers, & Kearl, 1963) suggest that graphs are more effective than tables or text for communicating numeric information, while other researchers (Vernon, 1950) offer evidence that contradicts this assertion. More likely, instructional effectiveness would result from a combination of instructional modalities and strategies (Kourilsky & Wittrock, 1987) and appropriate use of graphic design principles (Felker, 1980; Tufte, 1983). Some available research has been heavily criticized for poor experimental design or test validity (MacDonald-Ross, 1978) and a lack of a theoretical framework (Reynolds & Baker, 1987). Even so, there are some thoughtful guidelines for interpreting or constructing graphs available from several sources. Prominent among these are the texts of Cleveland (1985), Hartley (1995), Kosslyn (1994), Schmidt (1983), and Tufte (1983, 1990).

Teachers' Reports of Graphical Display Use

Both teachers and theorists posit that students understand graphs poorly. When Barkley (1987) presented 125 seventh and eighth graders with the simple graphing question shown in Figure 1, sixty percent chose answer B instead of the correct answer--A.

Jan walks away from a mark on the floor at a steady rate and then walks back toward it. Which distance graph below would best describe her walk?

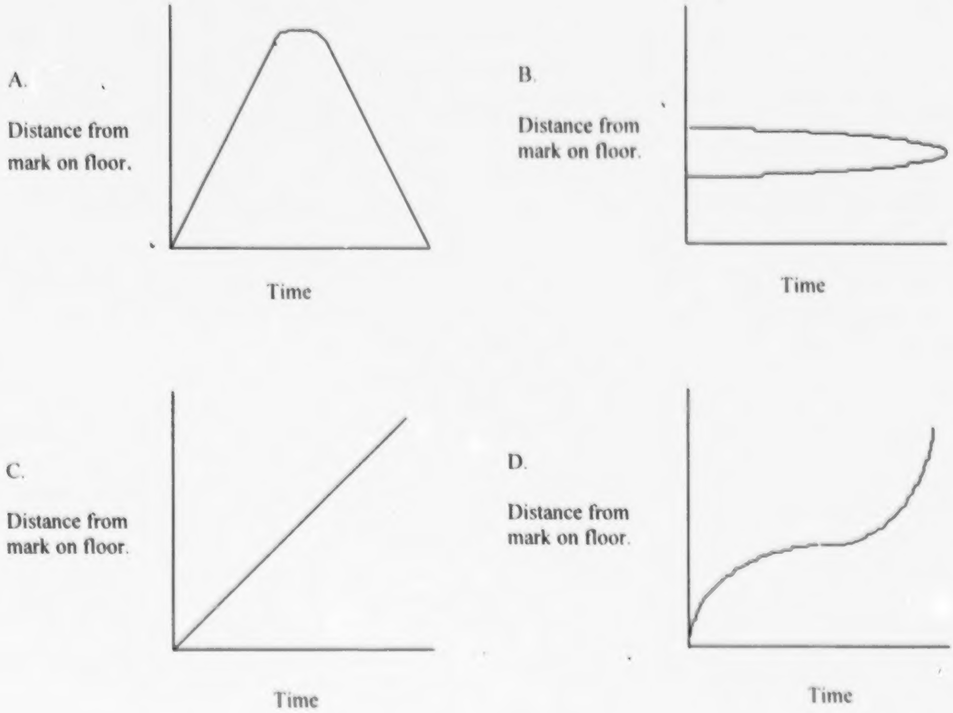


Figure 1. Adapted from Barkley, T. (1987, February). A graph is worth how many words? *Classroom Computer Learning*, p. 46.

This survey sampled K-12 teachers in a large Southern US school district regarding the use of quantitative graphical displays in public schools. Scant current information exists in the literature regarding chart and graph usage in the schools. Even more meager is the baseline information regarding that which students have had an opportunity to learn in school. Information about how teachers use graphical displays in varying subject content and in different grade levels is limited, although there are some materials available. For example, Brasell (1990) contends that in science areas, graphing is generally taught at the elementary level (p. 72). Although this type of anecdotal information, related national tests (NAEP, 1985), and state guidelines are somewhat illuminating, teachers have rarely been asked to honestly report information about graphical instruction and learning where their anonymity was protected.

The survey itself dealt with 12 major topics including use, familiarity, and interpretation problems. Teachers were queried about how they used charts, how they were trained to use charts, which charts they employed most frequently, and which charts they believed to be appropriate for student use. Teachers were also questioned about how students should use charts and what learning strategies students used to understand charts. They were asked when and in what content areas students should be introduced to charts. Finally, teachers were asked to describe how they taught students to analyze and construct charts. Because a pie chart, for example, is also called a circle graph, and to simplify our communication with teachers, we defined a "chart" or "graph" as any graphical display of quantitative data. We acknowledge that the present study is limited to teachers' reports of what occurs in schools.

Method

Subjects were 429 teachers from a large southeastern school system. The subjects taught kindergarten through 12th grade in a variety of subject matter. Sometimes an instructor would teach in as many as three different content areas. The schools in the system were grouped into elementary, middle, and high schools. Based on proportionate student population, seven schools were randomly selected from the elementary school list (including Kindergarten teachers, $n = 250$), four middle schools from the middle school list (grades 7 and 8, $n = 58$), and four high schools from the high school list ($n = 121$) of the system's schools. Due to some teachers' concerns about anonymity, age of the subjects, gender, and ethnic composition were not collected. According to system administrators, however, the schools represented a cross section of racial and rural/urban composition. In the fifteen schools from which the sample was drawn, 24% of the teachers were female, 76% were male, 28% were African-American, and 72% were white. Participation by teachers in this research was voluntary, however, as the data was collected during the teachers' regularly scheduled meeting or inservice periods and had the backing of school administrators, none of the teachers declined to participate. Approximately 54% of the teachers in the surveyed schools attended these in-services. Teachers were assured, in writing, that all data gathered would be completely anonymous. Most participants were experienced teachers (M years teaching = 12.7, $SD = 8.9$). The highest level of educational achievement of the teachers was a bachelor degree (51%), 48% had attained a masters degree, and 1% had a doctoral degree. Mean class size was 30, $SD = 22$.

Instrument and Procedures

A 78 item survey (including eight open-ended items) was designed to measure the use of quantitative displays in the schools. It was piloted with thirty experienced K-12 instructors and revised based on their comments. The instrument was administered on location, usually during teacher in-service meetings. Subjects were introduced to one of the experimenters, usually by the principal of the school. The experimenter gave a short description of the goals of the survey while a one-page information sheet and the survey itself were distributed. After the researchers answered questions, usually relating to the use of their responses, teachers took approximately twenty minutes to complete the survey.

Results[1]

Use of Charts in Teaching

A slight majority of teachers (55.3%) use graphical displays often or very often in their teaching. Thirty-eight percent of the teachers used charts sometimes. Only 6% reported not using charts at all. As Figure 2 indicates, there was a much more frequent use of charts by instructors in kindergarten and elementary grades than in junior high and senior high schools ($\chi^2 = 51, p < .001$).

---

[1] Summarized responses to the 78-item survey and the marginals for survey questions may be viewed at http://www.coe.usouthal.edu/techReports/notes.html (technical report #96-2)

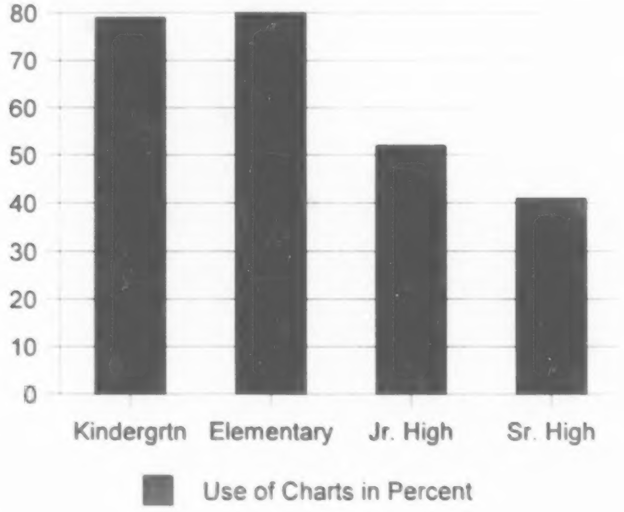


Figure 2. Use of charts by grade level.

Problems in Interpreting Charts

Teachers reported some common problems students had in correctly interpreting charts were: recognizing that chart intervals are scaled using standardized units (reported by 38.9%), recognizing patterns or trends in charts (35.7%), converting a point on the chart to a number (26.3%), and numerically comparing two different points on a chart (23.8%). See Table 1.

Table 1
Problems in Interpreting Charts

Do not recognize standardized units	38.9%
Do not recognize patterns (trends)	35.7%
Do not understand chart axis interval	34.5%
Convert chart point to number	26.3%
Cannot compare two points on chart	23.8%

Familiarity with Charts

About half (51.7%) of the instructors reported that they knew students had used charts before entering their class. Only 10.5% of the instructors reported that students had not used charts before entering their class. Instructors reported students were most familiar with bar charts (53.4%), followed respectively by pie charts (31.5%), line charts (29.84%), and combination charts (18.2%).

Ease of Use

Bar charts were reported to be easy or very easy for students to understand by 68.5% of instructors. Line charts, pie charts, and combination charts were reported to be easy or very easy for students to understand by 44.3%, 42.2%, and 23.3% respectively.

Charts Employed Most Frequently

Of the four types of charts considered in the survey, bar charts were reported to be used most frequently for instruction (78.8%); followed by line charts (52.0%); pie charts (48.7%); and lastly, combination charts, e.g., bar and line (37.5%).

Appropriateness for Student Use

Instructors recounted that bar charts were the most appropriate for student use (60.1%), followed by line charts (33.6%), pie charts (28.7%), and combination charts (35.9%).

Table 2
Charts Reported by Teachers to be Most Familiar, Used Most Often, and Most Appropriate for Students

	Most Familiar	Used Most Often	Most Appropriate
Bar	53.4%	78.8%	60.1%
Line	29.8%	52.0%	33.6%
Pie	31.5%	48.7%	28.7%
Combination	18.2%	37.5%	35.9%

Learning Strategies

Visualization was reported by instructors to be the most common strategy that students used to understand charts (69.5%), followed by demonstration and practice (54.5%), and concrete examples (47.1%). Only 11% of the instructors expressed that students used metaphors or analogies to understand charts.

In terms of effectiveness, visualization was considered effective or very effective by 79.1% of the instructors, concrete examples by 79.7%, and demonstration/practice by 78.8%. Metaphors and analogies were considered much less effective, with only 35.4% of the instructors considering that strategy to be effective or very effective.

Visualization, demonstration/practice, and concrete examples were considered helpful strategies for students to employ when using charts by 66.2%, 63.9%, and 66.2% of the instructors respectively. Again, metaphors and analogies were only considered helpful by 35.2% of the instructors. The use of any learning strategy to understand data presented in charts, however, declines as grade level increases.

Table 3
Learning Strategies Considered by Instructors to be Most Used by Students; Most Effective in General; and Most Effective for their Students to Use

	Strategies Used by Students	Most Effective Learning Strategy	Most Effective for Students
Visualization	69.5%	79.1%	66.2%
Concrete examples	47.1%	79.7%	66.2%
Metaphors/Analogies	11.1%	35.4%	35.2%
Demonstrations	54.5%	78.8%	63.9%

Combining Charts With Text

Greater than half (52.4%) of instructors reported students paid more attention to text combined with charts. Some instructors were unable to tell any difference (33.1%). Only a small percentage (7.7%) of instructors indicated students paid less attention to text with charts. See Figure 3.

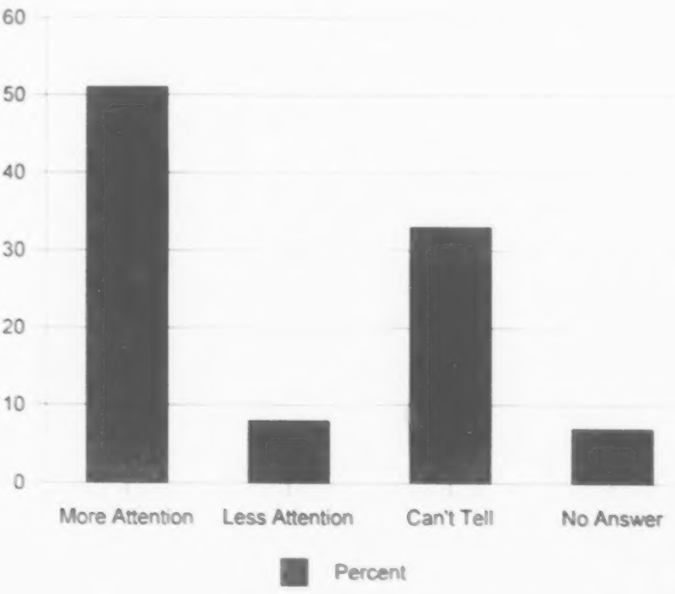


Figure 3. Teachers' perception of whether students pay more or less attention to text combined with graphical displays.

Teachers' Training in Chart Use

One-half (50.6%) of the respondents had some formal instruction about teaching students how to use charts. Instructors with a masters degree or greater were significantly more likely to have had formal training in interpreting or constructing graphs ($\chi^2 = 7.77, p < .01$).

Understand, Use or Construct?

More instructors suggested that it was more important for students, at the grade level they taught, to *understand* charts (48.5%), or **use** charts (42.4%) than it was for students to **construct** charts (28.7%). When teachers presented charts, 87% of the teachers reported they required students to interpret the information.

Teachers reported that students *constructed* graphs more frequently in Math (53.6%), and Science (35%), Social Studies (31.2%), English (15.9%), and History (14.9%). Foreign Language, Art and vocational areas reported a low incidence of student graph construction.

Academic Subjects Using Charts

The most frequent academic subjects in which charts were used for instruction were Math (52.2%), Social Studies (38.2%), Science (37.1%), English (27.5%), and History (19.6%). Areas where charts are not frequently used for instruction again include Foreign Language, Art, and vocational areas. See Figure 4.

When Should Students be Introduced to Charts?

Slightly less than half (48.5%) of the instructors reported students should be introduced to charts in Kindergarten. A lesser percentage (39.2%) of teachers would introduce children to charts in the first through third grades.

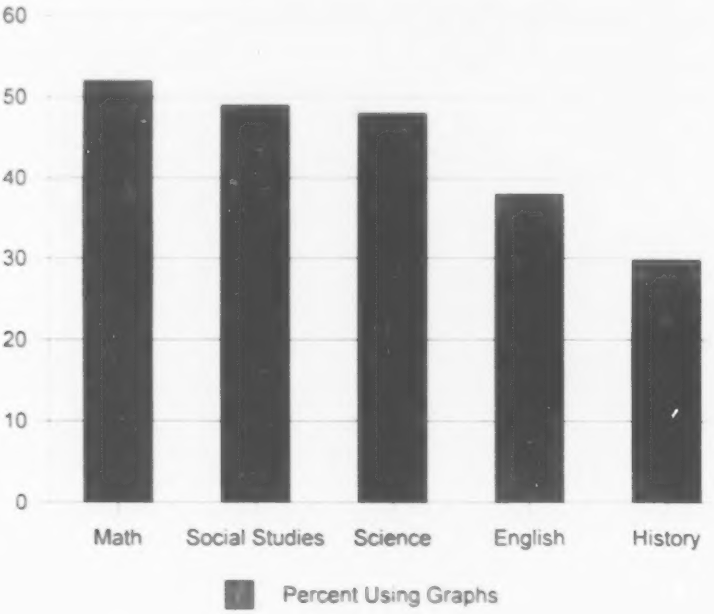


Figure 4. Use of graphical displays by academic subject.

How Do Teachers Teach Students to Read and Construct Charts?

Teachers were also asked to **describe** how they teach graphs. Of those responding, 25.6% mentioned using examples that related to their students and their daily lives. Other commonly mentioned methods included students constructing their own charts (24.8%), demonstration of charts (10.7%), and modeling (9.9%). Other methods described included using questions, discussion, and visualization.

Some responses suggested a clever use of the charting process. One example is shown in the following quotation from an elementary school teacher.

> M & M charting is a favorite of mine. Charting for the colors and also to determine more, less, most and least. Then the best part, they get to eat the M & Ms. This is a charting experience my students love to construct and one they also do at any holiday. I use the same principle as above using holiday candy in charting similar shapes.

Another inventive (and less caloric) approach to teaching the charting process was reported by an instructor who emphasized the motivational component of relevance.

> I had my math kids make a bar chart using information on the class (i.e., how they breakfasted, showered, petted a dog, watched TV, etc.) Then as a class we changed the bar chart to a picture chart, line chart, pie chart, and any other kind we tried to learn. They enjoyed tallying the data and interpreting it because it was about them. It was a very gratifying experience.

Discussion

A major trend reported in this study was that most instruction in graphical display use occurs in kindergarten and elementary schools. Given the more active nature of instruction at that level, it is not surprising that the attempt to make data more concrete is more common. It may be inferred from the data that many teachers think instruction in and use of graphical displays is less important past that point because older students have more well developed reading skills. The data presented in this study indicate that use of graphical displays drops to almost half in junior and senior high schools. Similarly, the use of any learning strategy to understand graphical displays declines as grade level increases. The corollary here is that referential connections (links between verbal and nonverbal symbolic systems) are assumed by teachers to be increasingly unimportant as students progress in school. This assumption contradicts recent psychological theories especially those of dual-coding (Paivio, 1986a), conjoint retention (Robinson, Katayama, Fan, 1996), and the contiguity effect (Mayer & Gallini, 1990).

According to dual-coding theory, activating an imagery system (such as graphs) can unify multiple objects into an integrated image (Clark & Paivio, 1991). Such an integration can facilitate memory for textbooks and other school materials. Paivio (1971, 1986) holds that there are three variables which increase the probability of imagery processing. These are: (1) instructions and related context effects, (2) concreteness, and (3) individual differences of learners. Consider these variables in relation to the findings of the present study. First, Paivio and his associates assert that students are more likely to generate mental images if instructed to do so than left to their own devices. The present study would suggest that as students progress through school and increase the use of verbal systems they receive instructions to use visual systems such as graphs less often (see Figure 1). The second determinant of imagery processing is concreteness or imagery value. This study found that graphs are being used most often in the highly quantitative areas such as Mathematics and Science. Graphs could also be employed for a variety of purposes in subject areas such as History including categorization and sorting, comparison and contrast, similarities and trends, summarization, and so forth. Graphs are a way of emphasizing concrete phenomena over the abstract. The implications of Paivio's third determinant, individual differences, are also pertinent. According to Clark and Paivio (1991), students who have trouble using image systems may fail to remember texts that benefit from imaging, may not understand geography of other spatial facts in a concrete fashion, and may do poorly in other areas such as visualizing steps of geometric proofs or spelling difficult words (p. 157). In this light, the sharp decline in employing graphical displays in instructional activities after elementary school seems foolish.

The data also suggest that bar charts are used much more frequently than other forms of graphical displays. There has long been evidence in the literature that bar charts are effective for comparisons (Croxton & Stein, 1932) and legibility (Culbertson & Powers, 1959). The heavy use of this type of chart in an age when other forms of graphical displays are readily available and may be more appropriate may be related to the limited use of graphs and charts in the formal academic and inservice training of junior and senior high school teachers. Approximately one-half of all of the teachers in this survey received no formal training in how to use graphical displays in their instructional activities. This would appear to be a major factor regarding teachers' lack of use or misuse of graphical displays in teaching. On the positive side, this survey suggests that formal training in interpreting or constructing graphical displays is significantly higher when teachers have attained a graduate degree.

One of the most common uses of graphical displays is with text. This use continues beyond school settings and is a mainstay in many adult communications (e.g., quarterly reports or newspaper accounts). The assumption is that graphs help to emphasize or explain more abstract data presented in a textual form. Noteworthy, therefore, is that only about half the teachers surveyed suggested that students paid more attention to text combined with charts. One explanation for this teacher perception may be that graphical displays are taught much more frequently in subjects that use less text (e.g., Mathematics) than in subjects that are heavily dependent on text (e.g., History). By contrast, a growing number of research studies suggest that student learning is improved by presenting text and graphical displays together (Glenberg & Langston, 1992; Purnell & Soloman, 1991; Waddil, McDaniel, & Einstein, 1988).

Some researchers suggest that constructing graphs (as opposed to reading or interpreting them) may increase the learning of graphic representations (Brasell, 1987). Data collected in this survey suggest that most teachers place less emphasis on constructing graphical displays. Most of the teachers who do encourage their students were at the elementary level. The exception to this trend was in mathematics, where newer technologies such as graphing calculators and computer programs may be making the process easier at junior and senior high school grades (Linn, Layman, & Nachmias, 1987).

Implications

Colleges of Education, teacher continuing education programs, and inservice administrators would do well to incorporate formal training experiences in using graphical displays. That one-half of the teachers in this survey reported receiving no training at all in the instructional use of graphical displays reflects poorly on these programs. In the upper grade levels, where graphical displays use is at its lowest level, such promising techniques as real-time graphing (Brasell, 1987) have great promise for allowing students the opportunity to construct graphical displays and aid in their comprehension of data.

Only about half of the teachers in this study reported that students paid more attention to graphs combined with text. If this perception is true, it could be because the instructional materials have failed to make a "visual argument" (MacDonald-Ross, 1978). Frequently, the cause of this is the failure by courseware developers to reach a harmony between graphic and instructional design principles. Graphical displays should embody information in a way that delivers a message to learners. When used with text they should use a design layout that tracks the graphical display to textual content.

Well-researched principles combine the best of both instructional and graphical design. For example, how information is "chunked" or summarized (Miller, 1956), influences the amount of human memory required for the display (Simcox 1983a; 1983b). Discriminating color use (Waller, Lefrere, & MacDonald-Ross, 1982) and related typographic cuing (Misanchuk, 1992) guide the learner's exploration of printed materials. Simplicity (Head & Moore, 1989), learner preference (Fisher, Dempsey, & Marousky, 1997), and graphical integrity (Tufte, 1983) may be used intentionally to clarify, gain attention, and promote retention.

Students should be encouraged to construct graphs more frequently in text-laden academic subjects. For instance, this survey indicated that graph construction in History is especially low. Graphing could certainly be a valuable tool for students to make historical trends more concrete or for a variety of other explanatory or exploratory purposes. Cross-curricula teacher training innovations modeled after the successful "Writing Across the Curriculum" program (Johnson, 1989), could assist instructors in incorporating graphical display activities into instruction. Allowing students to work in groups may encourage more successful graphical display construction (Jackson, Berger, & Edwards, 1989).

Educational researchers and instructional designers would be wise to study those teachers who are using clever strategies to teach students to construct or interpret graphical displays. Although there are a limited number of "how-to" articles available in teacher-oriented magazines (Paine, 1983), insufficient information is available to teacher educators about the effectiveness of imaginative methods which incorporate graphical displays into curricula. Anecdotal information in this survey found that some teachers have initiated or adopted some interesting techniques for making graphical displays more relevant to students' learning processes. By studying the instructional methods used by these teacher-innovators, qualitative researchers, in particular, have an unusually rich opportunity to contribute to the literature on using and understanding graphical displays of the complex information that permeates our lives.

References

Barkley, T. (1987, February). A graph is worth how many words? *Classroom Computer Learning*, 46-50.

Bertin, J. (1973). *Semiologie graphique* (2nd Ed.). The Hague: Mouton-Gautier. English translation by William Berg & Howard Wainer (1983) and published as the *Semiology of graphics*, Madison, WI: University of Wisconsin Press.

Brasell, H. M. (1990). Graphs, graphing, and graphers. In M.B. Rowe, (Ed.) *What Research Says to the Science Teacher*, Vol. 6., Washington, DC: National Science Teachers Association.

Brasell, H. M. (1987). The effect of real-time laboratory graphing on learning graphic representation of distance and velocity. *Journal of Research in Science Teaching*, *24*, 384-395.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*, 149-210.

Cleveland, W. S. (1985). *The elements of graphing data*. Montery, CA: Wadsworth.

Cleveland, W. S. (1984). Graphs in scientific publications. *The American Statistician*, *38*(4), 261-269.

Croston, F., & Stein, H. (1932). Graphic comparison by bars, squares, circles, and cubes. *Journal of the American Statistical Association*, *27*, 54-60.

Culbertson, H., & Powers, R. (1959). A study of graph comprehension difficulties. *Audio Visual Communications Review*, *19*, 399-416.

Feliciano, G. D., Powers, R. D., & Kearl, B. E. (1963). The presentation of statistical information. *AV Communication Review*, *11*, 32-39.

Felker, D. B. (Ed.). (1980). *Document design: A review of the relevant research*. Washington, DC: American Institute for Research.

Fisher, S. H., Dempsey, J. V., & Marousky, R. M. (1997). Data visualization: Preference and use of two-dimensional and three-dimensional graphs. *Social Science Computer Review*. *15*(3), 256-263.

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language*, *31*, 121-159.

Head, J., & Moore, D. (1989). The effect of graphic format and cognitive style on recall of quantitative data. *Canadian Journal of Educational Communication*, *20* (1), 3-15.

Hartley, J. (1995). *Designing instructional text* (2nd ed.). New York: Nichols.

Jackson, D. F., Berger, C. F., & Edwards, B. J. (1989, April). *The student as grapher: Microcomputer-assisted thinking skills*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Johnson, L. L. (1989). Learning across the curriculum with creative graphing. *Journal of Reading*, *32*, 509-518.

Kosslyn, S. M. (1994). *Elements of graph design*. New York: W.H. Freeman and Company.

Kourilsky, M. & Wittrock, M. C. (1987). Verbal and graphical strategies in the teaching of economics. *Teaching & Teacher Education*, *3*(1), 1-12.

Kulhavy, R. W., Lee, J. B., & Caterino, L. C. (1985). Conjoint retention of maps and related discourse. *Contemporary Educational Psychology*, *10*, 28-37.

Lin, C. Y. (1979). Imagery in matematical thinking and learning. *International Journal of Mathematics Education in Science, and Technology*, *10*(1), 107-111.

Linn, M. C., Layman, J. W., & Nachmias, R. (1987). Cognitive consequences of microcomputer-based laboratories: Graphing skills development. *Contemporary Educational Psychology*, *12*, 244-253.

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, *82*, 715-726.

MacDonald-Ross, M. (1978). Graphics in text. In Shulman, L.S. (Ed.). *Review of Research in Education* (Vol. 5). Itasca, IL: F.E. Peacock Publishers, Inc.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-96.

Misanchuk, E. R. (1992). *Preparing instructional text: Document design using desktop publishing*. Englewood Cliffs, NJ: Educational Technology Publications.

Moore, P. J. (1993). Metacognitive processing of diagrams, maps, and graphs. *Learning and Instruction*, *3*, 215-226.

NAEP. (1985). *National assessment of educational progress, reading in America: A perspective on two assessments* (Report 06-R-01) Washington, DC: United States Government Printing Office.

Paine, C. (1983, January). Graphing matters. *Learning*, 38-40.

Paivio, A. (1983) The empirical case for dual coding. In J.C. Yuille (Ed.). *Imagery, memory and cognition*. Hillsdale, NJ: Erlbaum.

Paivio, A. (1986a). *Mental representations: A dual coding approach*. New York: Oxford University Press.

Paivio, A. (1986b). Dual coding and episodic memory: Subjective and objective sources of memory trace components. In F. Klis & H. Hafgendorf (Eds.). *Human memory and cognitive capabilities: Mechanisms and performances* (Part A, pp. 225-236). Amsterdam: North-Holland.

Paivio, A. & Csapo, K. (1975). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology, 5*, 176-207.

Purnell, K. N., & Soloman, R. T. (1991). The influence of technical illustrations on students' comprehension in geography. *Reading Research Quarterly, 26*, 277-299.

Reynolds, R. E., & Baker, D. R. (1987). The utility of graphical representations in text: Some theoretical and empirical issues. *Journal of Research in Science Teaching, 24* (2), 161-173.

Robinson, D. H., Katayama, A. D., & Fan, A. C. (1996). Evidence for conjoint retention of information encoded from spatial adjunct displays. *Contemporary Educational Psychology, 21*, 221-239.

Schmidt, C. F. (1983). *Statistical graphics: Design principles and practices*. New York: John Wiley and Sons.

Simcox, W. (1983a). *A method for pragmatic communication in graphic displays*. Wellelesly, MA: Consulting Statisticians, Inc.

Simcox, W. (1983b). *Memorial consequences of display coding*. Wellelesly, MA: Consulting Statisticians, Inc.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1983). *The visual design of quantitative information*. Cheshire, CT: Graphics Press.

Vernon, M. D. (1950). The visual presentation of factual data. *British Journal of Educational Psychology, 20*, 174-185.

Waddil, P. J., McDaniel, M. A., & Einstein, G. O. (1988). Illustrations as adjuncts to prose: A test appropriate processing approach. *Journal of Educational Psychology, 80*, 457-464.

Waller, R., Lefrere, P., & MacDonald-Ross, M. (1982). Do you need that second color, *IEEE Transactions on Professional Communication, 25*(2), 80-85.

Wainer, H. (1992) Understanding graphs and tables. *Educational Researcher, 21*(1), 14-23.

Winn, W. (1991). Learning from maps and diagrams. *Educational Psychology Review, 3*, 211-247.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 43-52

Statistics Anxiety: A Function of Learning Style?

Anthony J. Onwuegbuzie
Valdosta State University

As the importance of research is being recognized increasingly, more teachers are required to enroll in research methodology courses as a necessary part of their graduate degree program. Unfortunately, it appears that these courses are exceedingly difficult for many students. Statistics anxiety appears to be a barrier to success in these courses. To date, no research has been conducted regarding the relationship between learning styles and statistics anxiety. Thus, the purpose of this study was to investigate this relationship in a research methodology course, using a multivariate analysis. Participants were 82 graduate students (90.2% teachers). A canonical correlation analysis yielded one statistically significant canonical root. In addition, two other canonical roots, although not statistically significant, appear to be educationally significant. In any case, the first canonical root suggests that classroom design, structure of the course, authority-orientation, auditory-orientation, food intake preference, time of day preference (i.e., morning vs. evening), and mobility preference, are related in varying degrees to worth of statistics, interpretation anxiety, test and class anxiety, computation self-concept, fear of asking for help, and fear of statistics teachers. Recommendations for future research are made, which include replicating the study using larger samples and different populations (e.g., undergraduate students).

Many researchers assert that an important way to facilitate school reform is to involve teachers in undertaking action research (Holly, 1991; Hovda & Kyle, 1984; McCutcheon, 1987; McKernan, 1987; Pine, 1986; Sardo-Brown, 1990). Action research is a continuous, self-reflective process which involves a critical examination of teaching practices or theories with a view to improving the quality of teaching as well as the education of students (McKernan, 1987). As such, research methodology courses, in which the techniques of action research are taught and emphasized, can empower teachers in their quest to improve teaching and learning (Clift, Veal, Johnson, & Holland, 1990; Hovda & Kyle, 1984). Consequently, in recent years, most graduate educational programs have required teachers to enroll in at least one research methodology course as a part of their degree program.

Unfortunately, anxiety induced by research methodology courses can be so great that undertaking these classes has come to be regarded by many as a negative experience (Onwuegbuzie, 1997). One reason for the negativity expressed by students stems from the fact that many who are enrolled in research methodology courses have had little or no formal exposure to statistics (Onwuegbuzie, 1997). Additionally, data indicate that many college students experience high levels of statistics anxiety when confronted with statistical ideas, problems, or issues, instructional situations, or evaluative situations (Feinberg & Halperin, 1978; Onwuegbuzie & Daley, 1996; Onwuegbuzie & Seaman, 1994; Roberts & Bilderback, 1980; Zeidner, 1991). Since statistical analyses typically are needed to address research questions and to test hypotheses, particularly in studies which utilize the quantitative paradigm, statistics anxiety can be a barrier to success in research methodology courses (Onwuegbuzie, 1997). Indeed, statistics anxiety has been found not only to be prevalent in research methodology courses, but also to affect a student's ability to acquire the skills, knowledge, and strategies necessary to interpret and to critique research reports, as well as to propose, to design, and to implement research studies (Onwuegbuzie, 1997).

The debilitative effects of statistics anxiety on statistics achievement (Benson, 1989; Onwuegbuzie & Daley, 1996; Onwuegbuzie & Seaman, 1995; Zeidner, 1991) and performance in research methodology courses (Onwuegbuzie, 1997) have been documented. Thus, statistics anxiety is similar to other types of academic-related anxiety, such as test anxiety (Galassi, Frierson, & Sharer, 1981; Hill, 1984; Lusk, 1983; Tobias, 1985; Wine, 1980), library anxiety (Onwuegbuzie, 1997), composition anxiety (Aldrich, 1982; Fox, 1980; Onwuegbuzie, 1997), and foreign language anxiety (MacIntyre & Gardner, 1994; Onwuegbuzie, Bailey, &

Anthony J. Onwuegbuzie is Assistant Professor in the Department of Educational Leadership at Valdosta State University, Valdosta, Georgia, where he teaches quantitative and qualitative research methodology courses, as well as intermediate- and advanced-level graduate statistics courses. Correspondence should be addressed to Anthony J. Onwuegbuzie, Department of Educational Leadership, College of Education, Valdosta State University, Valdosta, Georgia, 31698 or by e-mail at Tonwuegb@valdosta.edu

Daley, 1997; Phillips, 1992), in that it impedes academic performance.

A few correlates of statistics anxiety have been investigated in the literature. Specifically, females have been found to report higher levels of statistics anxiety than do males (Benson, 1989). In addition, Benson (1989) found a statistically significant negative correlation between number of college mathematics courses completed and mathematics self-concept and statistics anxiety. Tomazic and Katz (1988) suggested that academic major, academic status, perception of previous success in mathematics courses, and the time elapsed since students' last mathematics course were predictors of statistics anxiety. Finally, Roberts and Saxe (1982) found statistically significant correlations between statistics anxiety and basic mathematics skills, prior knowledge of statistics, statistics course grade, number of prior mathematics courses completed, the status of the course (i.e., required or elective), attitudes toward calculators, course and instructor evaluations, satisfaction with the statistics course, and gender.

Unfortunately, limited research exists on the characteristics of students with high levels of statistics anxiety (Auzmendi, 1991). Nevertheless, Onwuegbuzie, DaRos, and Ryan (1997) found that students with high levels of anxiety frequently reported that statistics is far removed from their field and that, consequently, they have difficulty adjusting their processing style to the study of statistics. Thus, although not yet tested empirically, learning style may be an antecedent of statistics anxiety. That is, college students' level of statistics anxiety may be moderated through their learning modality.

A number of studies have investigated the relationship between learning styles and achievement, in an attempt to identify the correlates of academic success. In the area of statistics, Elmore and Vasu (1986) found spatial visualization ability to be a modest predictor of statistics achievement, explaining 4.1% of the variance. Furthermore, Hudak and Anderson (1990) reported that formal operational ability and learning style also are predictors of statistics achievement. Specifically, these authors found that statistics achievement was related to the presence of the capacity to act as a formal operator and to the absence of a reliance on the concrete experiences learning style. Unfortunately, no correlation coefficients were reported. Similarly, Reece and Todd (1989) found that students who expressed a preference for the analyst style of thinking tended to have higher levels of performance on a test of statistical concepts than did those who did not express a preference for this style of thinking ($r = .28$).

Although only a few studies have investigated the relationship between learning style and anxiety, those which have documented a statistical association. Specifically, in a study of Navajo Middle school students, Hadfield, Martin, and Wooden (1992) found that spatial skills, discriminatory skills, and persistence orientation were negatively (statistically) related to mathematics anxiety. Unfortunately, again, no correlation coefficients were reported. At the college level, Reece and Todd (1989) observed that expressed preference for the formal-deductive style of thinking (i.e., synthesists and analysts) and mathematics anxiety are negatively (statistically) correlated. Specifically, a statistically significant negative relationship was reported between mathematics anxiety and analyst style of thinking ($r = -.38$) and between mathematics anxiety and synthesist style of thinking ($r = -.31$). Finally, McCoy (1992) found that the tactile/kinesthetic learning style was a significant predictor of mathematics anxiety. Interestingly, matching instruction to identified learning style (Lenehan, Dunn, Ingham, & Signer, 1994) or grouping students with peers who perceive and process materials in different ways (Price, 1991), appears to decrease levels of situation-specific anxiety.

An extensive review of the literature revealed no study which has investigated the relationship between college students' learning style and their level of statistics anxiety. Thus, this study was designed in order to identify a combination of learning modalities which might be correlated with a combination of statistics anxiety measures, using canonical correlation analyses. Canonical correlation is a statistical technique that breaks down the association between two sets of variables and is appropriate for describing the number and nature of canonical roots (Stevens, 1986). It was hoped that, through the application of canonical analysis, specific learning styles would be identified that might better explain the relationship between statistics anxiety and performance in research methodology courses. This, in turn, could assist in designing instructional strategies to improve any related deficiencies.

Method

Subjects

Graduate students enrolled in a College of Education research methodology course served as subjects. All 82 eligible students agreed to participate in the study (i.e., no student declined to participate). Of these, 70 (85.4%) were female and 74 (90.2%) were teachers in a public school system. Participants ranged in age from 22 to 56 (*mean* = 31.8, *SD* = 8.4). With regard to ethnicity, 72 (87.8%) were Caucasian, whereas 10 (12.2%) were African-American. Data were collected during two consecutive academic terms. Participation was voluntary and anonymous. All surveys were coded using student identification numbers in order to guarantee confidentiality.

Instruments

Instruments were administered at approximately the midpoint of the course, just prior to the students' midterm examination, since this is typically a time in which levels of statistics anxiety reach their peak (Onwuegbuzie et al., 1997). The following instruments were used in the study: the Statistical Anxiety Rating Scale (STARS) and the Productivity Environmental Preference Survey (PEPS).

The STARS was developed by Cruise and Wilkins (1980). STARS is a 51-item, 5-point Likert-format instrument which assesses statistics anxiety in a wide variety of academic situations. Although a few measures of statistics anxiety have been developed (e.g., Pretorius & Norman, 1992; Wilson, 1997; Zeidner, 1991), currently, STARS is the only multidimensional measure of statistics anxiety for which normative, reliability, and validity data have been reported. Using a multidimensional scale of statistics anxiety is justified further by Onwuegbuzie et al.'s (1997) in-depth qualitative study, in which statistics anxiety was found to be a multidimensional phenomenon. The STARS has six subscales, namely, *worth of statistics*, *interpretation anxiety*, *test and class anxiety*, *computation self-concept*, *fear of asking for help*, and *fear of the statistics instructor*. According to its authors, *worth of statistics* refers to a student's perception of the relevance of statistics. *Interpretation anxiety* is concerned with the anxiety experienced when a student is faced with making a decision from or interpreting statistical data. *Test and class anxiety* refers to the anxiety involved when taking a statistics class or test. *Computation self-concept* involves the anxiety experienced when attempting to solve mathematical problems, as well as the student's perception of her/his ability to do mathematics. *Fear of asking for help* measures the anxiety experienced when asking a fellow student or professor for help in understanding the material covered in class or any type of statistical data, such as an article or a printout. Finally, *fear of statistics teachers* is concerned with the student's perception of the statistics instructor. A high score on any subscale represents high anxiety in this area. Normative data have been gathered for this instrument. Cruise, Cash, and Bolton (1985) reported evidence of construct validity based on a factor analysis using 1,150 subjects in which six specific factors were identified after a varimax rotation. Loadings for these factors ranged from .48 to .86. Reliability of these factors, as measured by coefficient alpha, ranged from .68 to .94 (median = .88). In addition, Cruise et al. (1985), using a sample of 161 students, reported five-week test-retest reliability coefficients for each factor, which ranged from .67 to .83 (median = .76). For the present study, the reliability of the STARS subscales, as measured by coefficient alpha, ranged from .85 (*worth of statistics*) to .89 (*fear of asking for help*) (median = .86).

The PEPS, designed by Dunn, Dunn, and Price (1991), is an instrument that surveys individuals' preferences in each of 20 different modalities. The PEPS was developed through factor analysis using orthogonal (varimax) rotations. It is a comprehensive approach to the identification of how adults prefer to function, learn, concentrate, and perform during educational or work activities in the following areas: (a) environment (i.e., sound, temperature, light, and design); (b) emotionality (e.g., motivation, responsibility, persistence, and the need for either structure or flexibility); (c) sociological preferences (i.e., peer orientation, authority orientation); and (d) physical needs (e.g., perceptual preferences(s), time of day, intake, and mobility). Specifically, the PEPS measures preferences pertaining to the following 20 modalities: noise, light, temperature, design, motivation, persistence, responsibility, structure, peer orientation, authority orientation, multiple perceptual preferences, auditory, visual, tactile, kinesthetic, intake, evening/morning, late morning, afternoon, and mobility. Each subscale represents a learning modality. Performance on each of the 20 subscales is expressed in standard score units, which range from 20 to 80, with a mean of 50 and a standard deviation of 10. According to the instrument developers, individuals having a standard score of 40 or less or 60 or more find that modality important when they study or work. Individuals scoring between 40 and 60 typically differ with respect to how much that variable is important to them. Thus, for example, a high score on the late morning subscale (i.e., 60 or more) indicates a strong preference for undertaking difficult tasks in the late morning, whereas a low score (i.e., 40 or less) indicates that the individual does not prefer to undertake difficult tasks at this time. The reliabilities of the PEPS subscales range from .44 to .87 (median = .78), with nearly all the reliabilities exceeding .70 (Dunn et al., 1991). For the present study, the following factors were used: design, persistence, structure, learning alone, authority orientation, auditory, visual, tactile, kinesthetic, evening/morning, intake, and mobility. A subset of learning modalities was used in order to keep the ratio of subjects to variables close to 5 to 1, which is the minimum recommended ratio in canonical correlation analyses for obtaining reasonably stable effect size estimates (Thompson, 1990a). Indeed, with respect to the three time of day preference subscales (i.e., late morning, afternoon, and evening/morning), for the sake of parsimony, and since all the graduate-level research methodology courses were taught in the evening, only the evening/morning subscale was used. According to the authors of the PEPS, a high score on this subscale (i.e., 60 or more) indicates a strong preference for undertaking difficult tasks in the morning, whereas a low score (i.e., 40 or less) indicates a strong preference for undertaking difficult tasks in the evening. Unfortunately, the reliabilities of the subscales used for the present study were not available since the PEPS was scored by its owners. Finally, in the present study, scores on the PEPS were analyzed as continuous variables, instead of partitioning them (e.g., trichomotizing the scores into preference vs. neutral vs. non-preference), since to categorize a continuous variable is "to reduce its variance and thus its possible correlation with other variables" (Kerlinger, 1986, p. 558). Indeed, Pedhazur (1982) asserted that "categorization leads to a loss of information, and consequently to a less sensitive analysis" (pp. 452-453).

Analysis

A canonical correlation analysis was conducted to identify a combination of learning modality dimensions which might be correlated with a combination of statistics anxiety dimensions. Canonical correlation analysis is recommended to examine the relationship between two sets of variables, wherein each set contains more than one variable (Cliff & Krus, 1976; Darlington, Weinberg, & Walberg, 1973; Thompson, 1980, 1984). Indeed, as pointed out by Knapp (1978), "virtually all of the commonly encountered tests of significance can be treated as special cases of canonical correlation analysis" (p. 410). That is, canonical correlation analysis can be utilized to undertake all the parametric tests which canonical correlation methods subsume as special cases, including regression, analysis of variance, analysis of covariance, and *t*-tests (Thompson, 1988).

In the present study, the six dimensions of statistics anxiety were treated as the dependent multivariate profile, whereas the 12 dimensions of learning modality were treated as the independent multivariate profile. The number of canonical roots which can be generated for a given dataset is equal to the number of variables in the smaller of the two variable sets. Thus, six canonical roots were generated.

In the present study, three types of canonical coefficients were computed, standardized canonical function coefficients, index coefficients, and structure coefficients (see for example Reynolds, Stanton, McLean, & Kaufman, 1989). Standardized canonical function coefficients are derived weights applied to each of the variables in a given set in order to obtain the composite variate used in the canonical correlation analysis. As such, standardized canonical function coefficients are analogous to factor pattern coefficients in factor analysis or to beta coefficients in a regression analysis (Arnold, 1996). Index coefficients are the correlations between a given variable (dimension) and the scores on the canonical composite (i.e., latent variable) in the set to which the variable (dimension) does not belong (Thompson, 1980). Structure coefficients are the correlations between a given variable (dimension) and the scores on the canonical composite (i.e., latent variable) in the set to which the variable (dimension) belongs (Thompson, 1980). Thus, structure coefficients indicate the degree of relationship of a given variable in the set with the canonical composite for the variable set. The square of the structure coefficient is the proportion of variance that the original variable shares linearly with the canonical variate.

Results

Table 1 presents the correlation matrix from which the canonical roots were generated. The strength of the relationship between the two sets of variables was assessed by examining the magnitude of the canonical correlation coefficients. These coefficients indicate the degree of relationship between the weighted learning modality variables and the weighted anxiety variables. In addition, the significance of the canonical roots was tested via the F-statistic based on Rao's approximation (Rao, 1952).

The canonical analysis revealed that all six canonical correlations combined were statistically significant ($p < .01$). However, when the first canonical root was excluded, the remaining five canonical roots combined were not statistically significant. Similarly, with the removal of the first and second canonical roots, the remaining canonical roots combined were not statistically significant. Indeed, further removal of canonical roots also produced statistically nonsignificant results. Together, these results suggest that the first canonical function was statistically significant, but all subsequent canonical roots were not statistically significant. However, since the calculated probabilities are sensitive to sample size, particular attention should be paid to the educational (practical) significance of the obtained results (Thompson, 1980). The educational significance of canonical correlations typically are assessed by examining their size (Thompson, 1980, 1984, 1988, 1990b). The canonical correlation indicates how much variance the sets of weighted original variables share with each other (Thompson, 1988). In the present study, the first canonical correlation ($R_{c1} = 69.8\%$), the second canonical correlation ($R_{c2} = 59.4\%$), and the third canonical correlation ($R_{c3} = 58.6\%$) appeared to be educationally significant, contributing 48.7% (i.e., R_{c1}^2), 35.3% (i.e., R_{c2}^2), and 34.3% (i.e., R_{c3}^2) of the shared variance, respectively. All subsequent canonical correlations each explained less than 20% of the variance. Consequently, only the first three canonical correlations were interpreted.

Table 1
Pearson Product-Moment Correlations of Perfectionism Dimensions and the Statistics Anxiety Dimensions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Worth of Statistics																	
2. Interpretation Anxiety	56																
3. Test and Class Anxiety	48	71															
4. Computation Self-concept	74	50	49														
5. Fear of Asking for Help	44	67	60	42													
6. Fear of Statistics Teachers	68	54	45	53	44												
7. Design	-19	-19	-13	-9	-7	-15											
8. Persistence	2	-1	7	7	-7	-14	-8										
9. Structure	8	30	24	10	20	16	25	-26									
10. Alone	21	17	4	22	-4	8	-11	8	-6								
11. Authority Orientation	6	15	16	4	-11	4	-7	12	10	46							
12. Auditory	-9	-23	-12	-3	-22	-10	-1	6	-11	8	10						
13. Visual	-0	2	1	8	16	6	-14	11	-1	-5	4	-40					
14. Tactile	21	-19	-16	-29	-9	-21	11	23	-1	4	9	-2	11				
15. Kinesthetic	2	7	9	4	0	-5	8	31	-1	17	10	7	-13	30			
16. Intake	15	19	18	4	12	7	-32	-16	-2	14	22	-19	7	-18	-6		
17. Evening/Morning	-20	-21	25	4	-7	-21	23	-10	5	-4	-7	-1	15	-6	-11	-3	
18. Mobility	29	22	27	19	4	20	-11	-16	12	27	28	11	-8	-5	13	31	-18

Note. Decimals omitted

Data pertaining to the first canonical root are presented in Table 2.

Table 2
Canonical Analysis of Learning Modality and Anxiety Variables
First Canonical Function

Variable (%)	Function	Index	Structure	Structure^2
Worth of Statistics	-.73*	-.33*	-.47*	22.09
Interpretation Anxiety	-.59*	-.47*	-.67*	44.89
Test and Class Anxiety	-.69*	-.49*	-.70*	49.00
Computation Self-concept	.81*	-.05	-.07	0.49
Fear of Asking for Help	.48*	-.17	-.24	5.76
Fear of Statistics Teachers	.17	-.22	-.32*	10.24
Adequacy (mean of structure^2)				22.08
Redundancy (Adequacy x R_{c1}^2)				10.75
Learning Modality				
Design	.01	.20	.30*	9.00
Persistence	-.05	-.05	-.08	0.64
Structure	-.33*	-.20	-.30*	9.00
Alone	.22	-.11	-.16	2.56
Authority Orientation	-.25	-.27	-.38*	14.44
Auditory	.33*	.14	.21	4.41
Visual	.24	.13	.19	3.61
Tactile	.19	.07	.10	1.00
Kinesthetic	-.07	-.09	-.13	1.69
Intake	-.10	-.24	-.34*	11.56
Evening/Morning	-.14	.41*	.59*	34.81
Mobility	-.24	-.32*	-.45*	20.25
Adequacy (mean of structure^2)				9.41
Redundancy (Adequacy x R_{c1}^2)				4.59

* loadings with large effect sizes

The redundancy estimates provide further insight into the relationship between the two sets of variables. The redundancy estimate is equal to the average of the squared multiple correlation of each of the variables in one set with all the variables in the other set (Pedhazur, 1982). The redundancy estimate (Table 2) indicates that, on average, 10.75% of the total variance in the set of anxiety components was accounted for by the linear combination of learning modalities, whereas 4.59% of the learning modality set variance was accounted for by a linear combination of the anxiety set. The adequacy estimate measures the degree to which each set's variance is represented in the canonical solution. The adequacy estimates in Table 2 indicate that 22.08% of the total anxiety set variance was represented in that set's canonical composite, and 9.41% of the learning modality set variance was represented in its composite. However, since redundancy coefficients are not truly multivariate, caution should be exercised in interpreting the redundancy coefficients, as recommended by Thompson (1988).

An examination of the standardized canonical function coefficients (Table 2) revealed that, using a cutoff correlation of 0.3 recommended by Lambert and Durand (1975) as an acceptable minimum loading value, five of the six statistics anxiety dimensions made an important contribution to the anxiety composite--with *computation self-concept* being the major contributor. Indeed, only *fear of the statistics teachers* did not appear to make an important contribution to this composite.

BEST COPY AVAILABLE

With respect to the learning modalities set, *structure* and *auditory* appeared to be the only major contributors. Interestingly, not all the standardized canonical function coefficients pertaining to the statistics anxiety dimensions were in the same direction. This was probably attributable to the fact that all intercorrelations involving these dimensions were moderate to large (see Table 1), suggesting that multicollinearity may be present. Indeed, standardized function coefficients typically are highly affected by the collinearity of the variables in a given set (Thompson, 1990b). Accordingly, structure coefficients always should be interpreted. These coefficients are particularly useful for assessing the nature of the relationships between two sets of variables (Thompson, 1984, 1988, 1990b).

The structure coefficients revealed that the following four statistics anxiety components made important contributions to the first canonical variate, respectively: *test and class anxiety*, *interpretation anxiety*, *worth of statistics*, and *fear of statistics teachers*. With regard to the learning modality cluster, *evening/morning* preference made the biggest contribution, with *mobility* making a moderate contribution, and *design*, *structure*, and *authority orientation* making modest contributions.

The index coefficients (Table 2) suggest that *test and class anxiety*, *interpretation anxiety*, and *worth of statistics* of the statistics anxiety cluster, and *evening/morning* and *mobility* of the learning modality set, appear to make significant contributions to shared variance.

The three canonical coefficients (i.e., standardized canonical function coefficients, index coefficients, and structure coefficients) pertaining to the first canonical root suggest that *classroom design*, *structure of the course*, *authority orientation*, *auditory orientation*, *food intake preference*, *evening/morning preference*, and *mobility preference*, are related to *worth of statistics*, *interpretation anxiety*, *test and class anxiety*, *computation self-concept*, *fear of asking for help*, and *fear of statistics teachers*, to varying degrees.

Data pertaining to the second canonical root are presented in Table 3. The standardized canonical function coefficients revealed that four of the six statistics anxiety dimensions made an important contribution to the anxiety composite. These dimensions were *worth of statistics*, *computation self-concept*, *fear of asking for help*, and *fear of statistics teachers*. *Computation self-concept* made by far the largest contribution. With respect to the learning modalities composite, *alone*, *tactile*, *intake*, *evening/morning*, and *mobility* appeared to be important contributors. As with the first canonical root, not all the standardized canonical function coefficients pertaining to the statistics anxiety dimensions were in the same direction, suggesting multicollinearity.

Table 3
Canonical Analysis of Learning Modality and Anxiety Variables
Second Canonical Function

Variable (%)	Function	Index	Structure	Structure^2
Statistics Anxiety:				
Worth of Statistics	-.26*	.35*	.59*	34.81
Interpretation Anxiety	-.11	.16	.27	7.29
Test and Class Anxiety	.12	.20	.34*	11.56
Computation Self-concept	.99*	.51*	.86*	73.96
Fear of Asking for Help	-.48*	.03	.04	0.16
Fear of Statistics Teachers	.49*	.37*	.63*	39.69
Adequacy (mean of structure^2)			27.91	
Redundancy (Adequacy x R_c^2)			9.83	
Learning Modality:				
Design	-.14	-.07	.12	1.44
Persistence	.14	.04	.07	0.49
Structure	.05	.06	.10	1.00
Alone	.46*	.21	.35*	12.25
Authority Orientation	.04	.10	.16	2.56
Auditory	.14	.06	.11	1.21
Visual	.15	.03	.05	0.25
Tactile	-.53*	-.30*	-.50*	25.00
Kinesthetic	.10	.02	.03	0.09
Intake	-.31*	-.02	-.04	0.16
Evening/Morning	-.35*	.02	.03	0.09
Mobility	-.34*	.20	.34*	11.56
Adequacy (mean of structure^2)				4.68
Redundancy (Adequacy x R_c^2)				1.65

* loadings with large effect sizes

The structure coefficients (Table 3) revealed that the following four statistics anxiety dimensions made important contributions to the second canonical variate, *worth of statistics*, *test and class anxiety*, *computation self-concept*, and *fear of statistics teachers*. With respect to the learning modality cluster, *tactile* made the biggest contribution, with *alone* and *mobility* making moderate contributions.

The index coefficients (Table 3) suggest that *worth of statistics*, *computation self-concept*, and *fear of statistics teachers* of the statistics anxiety cluster, and *tactile* of the learning modality set, appear to make significant contributions to shared variance.

The standardized canonical function coefficients, index coefficients, and structure coefficients pertaining to the second canonical root suggest that *worth of statistics*, *test and class anxiety*, *computation self-concept*, *fear of asking for help*, and *fear of statistics teachers* are associated with *alone*, *tactile*, *intake*, *evening/morning*, and *mobility*, to varying degrees. However, the second canonical root should be interpreted with caution since it was not statistically significant.

Table 4 presents data pertaining to the third canonical root. The standardized canonical function coefficients revealed that all statistics anxiety dimensions made an important contribution to the anxiety composite, with *interpretation anxiety* and *fear of asking* for help making the largest contributions. With respect to the learning modalities composite, *design, persistence, auditory, visual*, and *evening/morning* appeared to be important contributors. Again, the fact that the standardized canonical function coefficients of the statistics anxiety dimensions were not in the same direction suggests multicollinearity.

Table 4
Canonical Analysis of Learning Modality and Anxiety Variables: Third Canonical Function

Variable (%)	Function	Index	Structure	Structure^2
Worth of Statistics	-.33*	.15	.26	6.76
Interpretation Anxiety	-1.30*	-.06	-.11	1.21
Test and Class Anxiety	.54*	.17	.29	8.41
Computation Self-concept	-.46*	.02	.03	0.09
Fear of Asking for Help	83*	.28	47*	22.09
Fear of Statistics Teachers	54*	.26	.44*	19.36
Adequacy (mean of structure^2)				.65
Redundancy (Adequacy x R_{c1}^2)				3.32
Learning Modality:				
Design	30*	.02	.03	0.09
Persistence	-.87*	-.11	-.18	3.24
Structure	07	-.02	-.04	0.16
Alone	-.28	-.21	-.37*	13.69
Authority Orientation	-.21	-.18	-.31	9.61
Auditory	.30*	-.03	-.05	0.25
Visual	.54*	.12	.20	4.00
Tactile	-.14	.04	.07	0.49
Kinesthetic	.10	-.08	-.13	1.69
Intake	-.04	02	.03	0.09
Evening/Morning	-.51*	.11	.20	4.00
Mobility	.10	02	.03	0.09
Adequacy (mean of structure^2)				3.12
Redundancy (Adequacy x R_{c1}^2)				1.07

* loadings with large effect sizes

From the index coefficients, it can be seen that none of the statistics anxiety dimensions or learning modalities made important contributions to shared variance. With regard to the structure coefficients, *fear of asking for help*, and *fear of statistics teachers*, and *alone* were the only dimensions which made important contributions to the third canonical root.

The standardized canonical function coefficients, index coefficients, and structure coefficients pertaining to the third canonical root suggest that *worth of statistics, test and class anxiety, computation self-concept, fear of asking for help*, and *fear of statistics teachers* are associated with *alone, tactile, intake, evening/morning*, and *mobility*, to varying degrees. However, as with the second canonical root, the third canonical root should be interpreted with caution since it was not statistically significant.

Discussion

The fact that both statistics anxiety (Cruise et al., 1985; Onwuegbuzie et al., 1977; Zeidner, 1991) and learning modality preference (Dunn et al., 1991) appear to be multidimensional constructs justifies the use of multivariate analyses in order to understand better their relationship. The canonical correlation analysis yielded one statistically significant canonical root. In addition, two other canonical roots, although not statistically significant, appear to be educationally significant. Although the second and third roots should be interpreted with caution, it is clear that the first canonical root suggests a large relationship between learning modality preference and statistics anxiety.

Comparison of the results of the present study with previous research is difficult, since an extensive review of the literature failed to reveal similar studies comparing the amount of shared variance between anxiety and learning modality components. However, the relationship found between statistics anxiety and learning styles is consistent with previous research in which a statistically significant relationship between mathematics anxiety and learning styles has been found (Hadfield et al., 1992; McCoy, 1992; Reece & Todd, 1989).

A limitation of the present study is the relatively small sample used. These results, therefore, need to be replicated with larger samples and with different populations (e.g., undergraduate students). Furthermore, the relationship between statistics anxiety and the learning modalities that were excluded from the present analysis should be investigated. Also, more research is needed on how to accommodate different learning styles in teaching statistical concepts to teachers.

It should be pointed out that the STARS, as is the case for other current measures of statistics anxiety, measures levels of debilitative anxiety. Although consequences of situation-specific anxiety usually are negative and debilitating, there are occasions when they can be facilitating. The latter is often the result of dealing with anxiety in a positive manner. In this instance, anxiety can act as a motivator (Phillips, Martin, & Meyers, 1972). Indeed, Alpert and Haber (1960) separated facilitating effects of test anxiety (task-relevant responses) from debilitating effects (task-irrelevant responses). Thus, future research also should investigate whether learning styles are related to facilitating anxiety.

Nevertheless, to the extent that the present results are generalizable, an important implication of these findings

BEST COPY AVAILABLE

is that, when introducing statistical concepts, instructors of research methodology courses need to pay attention to teachers' learning styles. Recognizing that individuals differ in cognitive style and acting on those recognitions might be a significant first step in reducing statistics anxiety. Focusing on statistical content alone and not attempting to cater to different learning styles may have dire consequences with respect to achievement of teachers in research methodology courses, and, hence, to the teacher-as-researcher movement.

References

Aldrich, P. G. (1982). Adult writers: Some factors that interfere with effective writing. *Research in the Teaching of English, 16,* 298-300.

Alpert, R., & Haber, R. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology, 61,* 207-216.

Arnold, M. E. (1996, January). *The relationship of canonical correlation analysis to other parametric methods.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.

Auzmendi, E. (1991, April). *Factors related to attitude toward statistics: A study with a Spanish sample.* Paper presented at the annual meeting of the American Educational Research Association, Chicago: IL.

Benson, J. (1989). Structural components of statistical test anxiety in adults: An exploratory model. *Journal of Experimental Education, 57,* 247-261.

Cliff, N., & Krus, D. J. (1976). Interpretation of canonical analyses: Rotated vs. unrotated solutions. *Psychometrica, 41,* 35-42.

Clift, R., Veal, M. L., Johnson, M., & Holland, P. (1990). Restructuring teacher education through collaborative action research. *Journal of Teacher Education, 41*(2), 52-62.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985, August). *Development and validation of an instrument to measure statistical anxiety.* Paper presented at the annual meeting of the Statistical Education Section. Proceedings of the American Statistical Association.

Cruise, R. J., & Wilkins, E. M. (1980). *STARS: Statistical Anxiety Rating Scale.* Unpublished manuscript, Andrews University, Berrien Springs, MI.

Darlington, R. B., Weinberg, S. L., & Walberg, H. J. (1973). Canonical variate analysis and related techniques. *Review of Educational Research, 42,* 131-143.

Dunn, R., Dunn, K., & Price, G. (1991). *Productivity Environmental Preference Survey.* Lawrence, KS: Price Systems, Inc.

Elmore, P. B., & Vasu, E. S. (1986). A model of statistics achievement using spatial ability, feminist attitudes and mathematics-related variables as predictors. *Educational and Psychological Measurement, 46,* 215-222.

Feinberg, L., & Halperin, S. (1978). Affective and cognitive correlates of course performance in introductory statistics. *Journal of Experimental Education, 46*(4), 11-18.

Fox, R. F. (1980). Treatment of writing apprehension and its effect on composition. *Research in the Teaching of English, 14,* 39-49.

Galassi, J. P., Frierson, H. T., Jr., & Sharer, R. (1981). Behavior of high, moderate, and low test anxious students during an actual test situation. *Journal of Consulting and Clinical Psychology, 49,* 51-62.

Hadfield, O. D., Martin, J. V., & Wooden, S. (1992). Mathematics anxiety and learning style of the Navajo middle school student. *School Science and Mathematics, 92,* 171-176.

Hill, K. T. (1984). Debilitating motivation and testing: A major educational program, possible solutions, and policy applications. In R.E. Ames and C. Ames (Eds.), *Research on motivation in education, 1,* (pp. 245-274). New York: Academic Press.

Holly, P. (1991). Action research: The missing link in the creation of schools as centers of inquiry. In A. Lieberman & L. Miller (Eds.), *Staff development for education in the '90s* (pp. 133-157). New York: Teachers College Press.

Hovda, R. A., & Kyle, D. W. (1984). Action research: A professional development possibility. *Middle School Journal, 15*(13), 21-23.

Hudak, M. A., & Anderson, D. E. (1990). Formal operations and learning style predict success in statistics and computer science courses. *Teaching of Psychology, 17*(4), 231-234.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.

Knapp, T. R. (1978). Canonical correlation analysis: A general significance parametric significance testing system. *Psychological Bulletin, 85,* 410-416.

Lambert, Z. V. & Durand, R. M. (1975). Some precautions in using canonical analysis. *Journal of Market Research, XII,* 468-475.

Lenehan, M. C., Dunn, R., Ingham, J., & Signer, B. (1994). Effects of learning style intervention on college students' achievement, anxiety, anger, and curiosity. *Journal of College Student Development, 35,* 461-466.

Lusk, S. L. (1983). Interaction of test anxiety and locus of control on academic performance. *Psychological Reports, 53,* 639-644.

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning, 44,* 283-305.

McCoy, L. P. (1992). Correlates of mathematics anxiety. *Focus on Learning Problems in Mathematics, 14,* 51-57.

McCutcheon, G. (1987). Teacher's experience doing action research. *Peabody Journal of Education, 64,* 116-127.

McKernan, J. (1987). Action research and curriculum development. *Peabody Journal of Education, 64*(2), 6-19.

Onwuegbuzie, A. J. (1997). Writing a research proposal: The role of library anxiety, statistics anxiety, and composition anxiety. *Library and Information Science Research, 19,* 5-33.

Onwuegbuzie, A. J., Bailey, P., & Daley, C. E. (1997). *Factors associated with foreign language anxiety.* Manuscript submitted for publication.

Onwuegbuzie, A. J., & Daley, C. E. (1996). The relative contributions of examination-taking coping strategies and study coping strategies on test anxiety: A concurrent analysis. *Cognitive Therapy and Research, 20,* 287-303.

Onwuegbuzie, A. J., DaRos, D., & Ryan, J. (1997). The components of statistics anxiety: A phenomenological study. *Focus on Learning Problems in mathematics, 19*(4), 11-35.

Onwuegbuzie, A. J., & Seaman, M. (1994). The effect of time and anxiety on statistics achievement. *Journal of Experimental Psychology, 63,* 115-124.

Pedhazur, E. J. (1982). *Multiple Regression in behavior research.* Fort Worth, TX: Holt, Rinehart and Winston.

Phillips, E. M. (1992). The effects of language anxiety on student oral test performance and attitudes. *The Modern Language Journal, 76,* 14-26.

Phillips, B. N., Martin, R. P., & Meyers, J. (1972). Interventions in relation to anxiety in school. In C. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press.

Pine, G. (1986). Collaborative action research and staff development in the middle school. *Middle School Journal, 18,* 33-35.

Pretorius, T. B., & Norman, A. M. (1992). Psychometric data on the statistics anxiety scale for a sample of South African Students. *Educational and Psychological Measurement, 52,* 933-937.

Price, E. C. (1991, August). *Learning from the past as we aim for the future through identifying students' learning styles to improve teaching/learning experiences in college students.* Paper presented at the summer conference of the Association of Teacher Educators, Minot, ND.

Rao, C. R. (1952). *Advanced statistical methods in biometric research.* New York: Wiley.

Reece, C. C., & Todd, R. F. (1989, November). *Math anxiety, attainment of statistical concepts, and expressed preference for a formal-deductive cognitive style among beginning students of research.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR.

Reynolds, C. R., Stanton, H. C., McLean, J. E., & Kaufman, A. S. (1989). The canonical relationship between the WISC-R verbal and performance scales. *Measurement and Evaluation in Counseling and Development, 22,* 69-72.

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and validity of a statistics attitude survey. *Educational and Psychological Measurement, 40,* 235-238.

Roberts, D. M., & Saxe, J. E. (1982). Validity of a statistics attitude survey: A follow up study. *Educational and Psychological Measurement, 42,* 907-912.

Sardo-Brown, D. (1990). Middle level teachers' perceptions of action research. *Middle School Journal, 22*(3), 30-32.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1980, April). *Canonical correlation: Recent extensions for modeling educational processes.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretations.* Newbury Park, CA: Sage Publications. (ERIC Document Reproduction Service No. ED 199 269)

Thompson, B. (1988, April). *Canonical correlation analysis: An explanation with comments on correct practice.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B. (1990a). Finding a correction for the sampling error in multivariate measures of relationships: A Monte Carlo study. *Educational and Psychological Measurement, 50,* 15-31.

Thompson, B. (1990b, April). *Variable importance in multiple regression and canonical correlation.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 317 615)

Tobias, S. (1985). Test anxiety: interference, defective skills and cognitive capacity. *Educational Psychologist, 3*, 135-142.

Tomazic, T. J., & Katz, B. M. (1988, August). *Statistical anxiety in introductory applied statistics.* Paper presented at the annual meeting of the American Statistical Association, New Orleans, LA.

Wilson, V. (1997, November). *Factors related to anxiety in the graduate statistics classroom.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN.

Wine, J. (1980). Cognitive-attentional theory of test anxiety. In I.G. Sarason (Ed.), *Test anxiety: Theory, research and applications* (pp. 349-385). Hillsdale, NJ: Erlbaum.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students - some interesting parallels. *British Journal of Educational Psychology, 61*, 319-328.

Footnote

I would like to express my gratitude to Christine E. Daley for her editorial assistance and to the four anonymous reviewers for their extremely helpful comments on an earlier version of this article.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 53-60

Topic Coverage in Statistics Courses: A National Delphi Study

Kathleen Cage Mittag
University of Texas at San Antonio

Elizabeth M. Eltinge
Texas A&M University

Statistics courses are taught in many different departments by people of different training. With the diversity of teaching settings for statistics courses, it is important that the professional body of statisticians come to some agreement of what should or should not be included in an introductory statistics course. The present study was an effort to contact peer-identified experts in statistics education and to provide them with a forum to reach consensus on what topics should be included in introductory statistics courses. Other issues of interest included order of presentation of topics and the percent of time devoted to each topic. A Delphi procedure was used as the research tool in the study.

There is a great deal of interest in statistics education, both at the national level (Becker, 1996; Cobb, 1993) and the international level (Vere-Jones, 1995). It is becoming increasingly recognized that statistical literacy is an important component of education and many efforts are being made to enhance the learning experiences for students of introductory statistics courses (Cobb, 1993; Garfield, 1995; Gnanadesikan, Scheaffer, Watkins, & Witmer, 1997). The field of statistics education is relatively new and fast growing. Most of the literature to date addresses recommendations for instruction based on experiences and intuitions of individual instructors rather than on empirical studies (Becker, 1996).

One interesting problem facing statistics education is the diversity of settings for the teaching of introductory statistics courses. Various settings include departments of statistics, mathematics, psychology, engineering, business, and education, both at the junior college and the four year college or university level (Cobb, 1993; Oathout, 1995). With so many different settings it is important that the professional body of statisticians come to some agreement of what should or should not be included in an introductory statistics course. The present study was an effort to contact peer-identified experts in statistics education, and to provide them with a forum to reach consensus on what topics should be included in an introductory, algebra-based statistics course. Other related works offer guidelines for the content of statistics courses designed for specific groups of students, such as engineering majors (Hogg, 1985), mathematics majors (Alo, 1983; Gaudard & Hahn, 1991), and pre-medical majors (Grady, Looney, & Steiner, 1994). In the present study we desired to generate an essential list of topics for an introductory level statistics course to be taken by students from many different majors. We wished to utilize input from a number of recognized leaders in the field of statistics education. To this end we used a Delphi technique with three rounds of questionnaires.

The Delphi technique allows for individuals to express their beliefs and to prioritize these beliefs through a set of sequential questionnaires (Borg & Gall, 1983). Each individual involved in the survey is allowed to provide an equal voice in determining the outcomes of the investigation. Some advantages of the technique are: (a) it allows for critical reflection on topics identified, (b) the method is designed to allow for changing thoughts and priorities based on the ratings of others, and (c) it allows for anonymity of responses (Miller & Johnson, 1992).

Jackson (1991) conducted a Delphi study to generate a list of topics for a high school statistics course. At the time the present study was initiated the Jackson work was unpublished. We were, however, able to obtain copies of the questionnaire used and partial results. The Jackson study was useful in the development of this project; however, the present study differed substantially from the Jackson study in the method of panel selection and in the targeted course.

Kathleen Cage Mittag is Assistant Professor in the Mathematics/Statistics Division and Education Division, University of Texas at San Antonio. Elizabeth M. Eltinge is Adjunct Assistant Professor in the Statistics Department, Texas A&M University. The study was supported by a grant from the Center for Statistics Education, Texas A&M University. Correspondence concerning this article should be addressed to Kathleen Cage Mittag, Mathematics/Statistics Division, University of Texas at San Antonio, 6900 Loop 1604 West, San Antonio, TX 78249 or e-mail address, kmittag@lonestar.utsa.edu.

Method

Panel Selection

During a Delphi study a panel of experts in a particular field is asked to reach a consensus on a list of topics. The composition of the panel determines the validity of the results, thus the formation of the panel of experts is a very important step in a Delphi study. There are basically two methods of panel selection: (a) have the primary investigator name the panel or (b) have a committee with knowledge of the field recommend a panel. The first method is considered the least effective because of possible bias on the part of the primary investigator (Somers, Baker, & Isbell, 1984). For the present study a variation of the second method was used. Panel members were nominated by a randomly selected group of members of the Section on Statistical Education of the American Statistical Association.

To generate the random sample of members of the Statistical Education section, a set of mailing labels of the section members was obtained from the American Statistical Association. The members were listed in alphabetical order within a state. International members were excluded from the study. A fifty percent systematic random sample was obtained from this mailing list which resulted in a sample of size 374. This sampling scheme allowed proportional representation across the United States. The alphabetical listing of members within states should introduce no discernible bias into the sample.

Each person in the sample was then mailed a nomination form and a cover letter explaining the purpose of the study. The letter requested the nominations of three people whom the respondent felt to be leaders in the field of statistics education. The respondent was free to include his or her name as one of the nominees. The response rate was 52% with 194 nomination forms returned. No follow-up strategies were employed to contact nonrespondents. Because the formation of the panel consisted of identifying the most frequently mentioned names and there were relatively small numbers of names mentioned very frequently with a large number of names mentioned only once, it was felt that the information from the nonrespondents would not significantly alter the conclusions.

The nomination procedure resulted in a list of 298 nominees; 68 people on the list were nominated by three or more people. A panel of size 25 was deemed optimal for such a study (Linstone & Turoff, 1979). Anticipating a certain amount of attrition, the 32 most commonly nominated names were identified. All 32 nominees were contacted by telephone, with only two declining to participate. This resulted in a panel of size 30. After the first round of questionnaires was returned by panel members, one panel member decided that it was not possible for him to continue the process, so the final panel consisted of 29 members. Demographic information for the panel is presented in Table 1. Most of the panel members were affiliated with a college or university and had teaching as their primary responsibility. Almost 90% of the panel members listed statistics as their primary discipline. A list of the panel members is presented in the Appendix.

Table 1
Demographic Data for Panel Members ($N = 29$)

Variable	Percentages
Type of organization where currently employed:	
College or University	77.4
Federal Agency	3.2
Industry	12.9
Research/Development	3.2
Other	3.2
Major responsibility at present position:	
Research	14.3
Evaluation	1.8
Teaching	37.5
Consulting	16.1
Management/Administrate	12.5
Primary Discipline	
Education	3.4
Mathematics	3.4
Statistics	89.7
Other	6.9
Feel this course should be a two semester course?	
Yes	64
No	36

Note: The average enrollment at the universities of the panel members was 19,385 students. The minimum was 1,250 and the maximum was 41,000.

Initial Questionnaire

The initial questionnaire was created using a variety of resource information to generate a list of topics. As mentioned above, the study by Jackson (1991) on topics to be included in a high school-level statistics course provided useful information in generating a possible list of topics to be included in a college-level statistics course. Additional studies that provided information used in the development of the first questionnaire included: Allen, Efird, and Eliasziw (1990); Ames, Clason, and Urguhart (1990); Anderson and Loynes (1986); Cobb (1987); Foremen, Brown, and Behrens (1992); Grosof and Sardy (1990); Hogg (1992); Moore (1985); and Neter (1989), as well as various introductory statistics textbooks.

Results

Topics

At three different times panel members were mailed a list of topics and asked to rate the importance of each topic in an introductory statistics course. Panel members were allowed to add any new topics as they saw fit. Second and third lists were generated based on results of responses to previous rounds. There was a 97% response rate on each of the three rounds. For each round, one different person did not return the questionnaire.

The first round questionnaire consisted of 79 subcategories which were grouped into 12 main categories. Panel members were asked to rank each of the subcategories as "very important," "important," "slightly important," or "unimportant." The following definitions for these terms were printed on each questionnaire:

Very Important -- A most relevant point; first-order priority; definite inclusion as a topic;

Important -- Is relevant; second-order priority; does not have to be fully developed;

Slightly Important -- Marginally relevant; third-order priority; could be discussed; and

Unimportant -- No relevance; no priority; should be dropped as an item to considered.

Space was allowed on the response form for any panel member to add additional topics, either within one of the 12 main categories or as a separate new category. Panel members were also encouraged to make comments and suggestions which were recorded and sent to the panelists when the next round was mailed. It was important for all panel members to see all other comments in order to reach a consensus. Scores for each subcategory were averaged across responses, and the subcategories were arranged in rank order. A cut-off score was determined for elimination of less important subcategories. Entire categories were eliminated at times. The cut-off score was determined by looking for natural breaks in the data and was different for each questionnaire round. Nineteen subcategories were eliminated during the first round. Ten new subcategories were suggested by panel members to be included in Questionnaire 2.

In the second round, nineteen subcategories were eliminated and two new subcategories were suggested. Six subcategories were omitted after round three. This resulted in 54 subcategories on the final list which was very close to the 50 subcategories deemed desirable at the beginning of the study. The final list of topics is presented in Table 2.

Table 2
Consensus List of Topics and Suggested Order for an Introductory Statistics Course

Category and Topic
CATEGORY 1: DATA COLLECTION
1. Importance of randomization
2. Sample vs. experiment as data source
3. Types of data
4. Measurement
5. Context of data (Sources of data, analytic/enumerative, observational)
CATEGORY 2: SUMMARIZING DATA -- GRAPHS
1. Bivariate plotting techniques (scatterplots, lines, enhancements, smoothing)
2. Regression lines
3. Box plots
4. Stem-and-Leaf plots
5. Frequency histograms
6. Relative frequency histograms
7. Frequency tables
CATEGORY 3: SUMMARIZING DATA -- NUMERICAL
1. Measures of dispersion (range, variance, SD, IQR)
2. Measures of center (mean, median)
3. Population parameters vs. sample statistics
4. Outliers
5. Shape of the data
6. Quartiles
CATEGORY 4: WHY STATISTICS? OF WHAT USE WILL IT BE TO ME?
1. Identify the goals of the course and the ground rules to be used to guide progress toward these goals
CATEGORY 5: PROBABILITY DISTRIBUTIONS
1. Normal distribution
2. Independent vs. dependent events
3. Expected value
4. Central limit theorem
5. Simulation of possible outcomes
6. Binomial distributions
CATEGORY 6: EXPERIMENTAL DESIGN
1. Variability
2. Randomization
3. Replication
4. Control and experimental groups
5. Response and explanatory variables
6. Bias
7. Confounding variables
CATEGORY 7: ESTIMATION
1. Effects of sample size
2. Interval estimation and confidence intervals
3. Sampling distributions of statistics
4. Point estimation

(table continues)

BEST COPY AVAILABLE

Table 2 (continued)

CATEGORY 8: HYPOTHESIS TESTING
1. Meaning of statistically significant
2. Introduction to inference
3. p-values
4. Type I and Type II errors
5. Two population tests of means
6. One population tests of proportion
7. Paired vs. non-paired t-tests
8. One population tests of means

CATEGORY 9: CORRELATION AND REGRESSION
1. Regression equation (interpreting)
2. Correlation vs. causation
3. Prediction
4. Residual analysis
5. Pearson's correlation coefficient
6. Standard error

CATEGORY 10: CATEGORICAL DATA ANALYSIS
1. Two way frequency tables
2. Chi-squared test for independence

CATEGORY 11: ANALYSIS OF VARIANCE
1. Purpose of analysis of variance

Scope and Sequence

In addition to ranking topics, panel members were asked to indicate the order in which the main categories should be presented and to indicate the percent of time which should be devoted to each topic. The data for these issues was obtained from the third and final round of questionnaires. Ranks were averaged across panel member responses to arrive at the final order. Recommended order of presentation and recommended time allocations for the topics are presented in Table 3.

Table 3
Percent of Time Recommended for Each Topic

Category		Mean Percent
1.	Data Collection	8%
2.	Summarizing Data -- Graphs	12%
3.	Summarizing Data -- Numerical	8%
4.	Why Statistics?	4%
5.	Probability	8%
6.	Experimental	8%
7.	Estimation	14%
8.	Hypothesis Testing	11%
9.	Correlation and Regression	11%
10.	Categorical Data Analysis	7%
11.	Analysis of Variance	5%

Instructional Strategies

Panel members were also asked to indicate the appropriate percent of time which should be devoted to three different teaching approaches. The three approaches were: probability-based approach, data-based approach, and computer-based approach. Definitions for these terms were provided on each questionnaire as follows:

> Probability-based approach -- The probability-based approach is the "traditional" method of teaching statistics. This approach starts out spending one-fourth to one-third of the class time on basic laws and axioms of probability which are then used to build the concepts of hypothesis testing;
>
> Data-based approach -- The data-based approach uses real-life data to demonstrate statistical concepts; and
>
> Computer-based approach -- The computer-based approach has the students use statistical computer packages to teach statistical concepts. This could include simulations as well as interpretation of the output.

It was recognized that these three approaches are not necessarily mutually exclusive, but respondents were asked to indicate a relative preferred emphasis among the three. Results of recommended relative emphasis in teaching approaches are presented in Table 4.

Table 4
Suggested Instructional Approaches

Probability-based Approach	13%
Data-based Approach	49%
Computer-based Approach	28%
Other Suggested Approaches	10%

*Other: student labs and projects, experimental design, case studies

Finally, panel members were presented with various classroom activities and asked to indicate the percent of time which should be devoted to them. The activities listed were: lecture, discussion, book examples, projects, research examples, student presentations, and other activities. Results of recommended classroom time allocation are presented in Table 5.

Table 5
Teaching Approaches Emphasis

Approach	Allocation percent
Lecture	37.3%
Discussion	21.7%
Book Example	10.9%
Projects	9.6%
Research Example	10.5%
Student Presentations	5.6%
Other	4.4%

Panel Comments

Several panelists felt constrained by limiting their answers to the options allowed on the questionnaire. Certain topics prompted lively discussion and debate as reflected in the selected comments shown below:

Regarding data and measurement:

- "Data production is very important in all statistics/applications."
- "Data analysis gives students hands-on experience and teaches exploratory, questioning attitudes towards data -- the opposite of running data through software and rejecting at the 5% level."
- "Too much time spend on formal probability without sufficient base of data analysis turns people off."
- "My biggest surprise is for the low enthusiasm for types of data."
- "Measurement integrity is critical to the ability to make valid inferences, and it is an area (a) that students think they know well and (b) usually don't."

Regarding probability:

- "Simulations make probability more concrete to many students."
- "Probability is a very hard subject. Students learn to do problems but don't understand them. It's a barrier, not an aid, to understanding statistics. (That mathematicians think otherwise only reminds us that mathematicians think differently than most people.)"
- "Put in only enough probability to support subsequent statistical tools and analyses."
- "Probability ideas underlie all of statistical inference."

Regarding estimation and hypothesis testing:

- "Estimation is basic to understanding and appreciating inference."
- "For me, interval estimation has become more important than either the point estimation or the hypothesis test."
- "I prefer confidence intervals to tests."
- "Like it or not, other people expect us to teach hypothesis testing. They are best done using computer packages, and packages report *p*-values."

Regarding specific statistical techniques:

- "General linear modeling is more important than analysis of variance, both heuristically and in practice."
- "Regression lines are a must since relationships are the ultimate thing of interest."
- "Logistic regression will become more important as technology improves."
- "Forget Analysis of Variance."
- "Too bad no one wants to pay attention to categorical data analysis."
- "Misuse of regression is widespread and this should be dwelt upon."
- "Students should be aware of nonparametrics."

Regarding the Central Limit Theorem:

- "The Central Limit Theorem is crucial!"
- "Putting the Central Limit Theorem in an intro class is like deciding to read a Dickens novel in the middle of a gourmet meal: the novel is a wonderful creation, but there's no quick way to do it justice, and no matter how much or little time you give it, you're detracting from the meal. Any restaurant that followed this practice would go broke and would deserve it."

Other general comments:

- "Holy cow! How much can we expect in a first course?"
- "For social science students, validity and reliability are important."
- "I am mostly interested in what students remember 5 years later. My assumption is that this is not only their first course in statistics -- but most likely their last course."

Discussion

Topics

The final list of topics (Table 2) resulted in a rather traditional course and closely matches the typical table of contents found in most introductory statistics textbooks (Ames, Clason, & Urguhart, 1990). Panelists noticed this tendency toward traditionalism, and some expressed concern:

- "I find that these categories make it very hard to escape from certain patterns of the traditional approach -- patterns which present statistics as a collection of tools and techniques with rules for when to use them, along with the departed ghost of an underlying mathematical theory."
- "The rounds have converged toward over much conservatism for me."

The list of topics in Table 2 matches the results from other related studies. Giesbrecht, Sell, Scialfa, Sandals, and Ehlers (1994) surveyed instructors of introductory statistics courses at the University of Calgary. Lopez and Mertens (1994) surveyed members of the Educational Research Special Interest Group of the American Educational Research Association. Giesbrecht (1996) combined the work of these two studies and concluded the most important topics to be included in an introductory statistics course are: summarizing data and graphs, summarizing numerical data, probability and probability distributions, estimation, hypothesis testing, correlation, and regression.

The one exception to a traditional list was Category 11 "Why statistics?" This category reflects the growing interest in incorporating certain quality control issues into the statistics curriculum. There was much discussion among panel members on the inclusion of this topic because it did not deal with content. However, by our rules of consensus voting, it was deemed important by the panel members and thus included as a category.

Scope and Sequence

The recommended order in which to present topics shown in Table 3 appears to be traditional. However, this order can be looked upon as a logical way in which to build concepts within the course. Other authors also suggest the use of data, graphics, and exploratory data analysis before the introduction of classical methods (Bradstreet, 1996). Several panelists expressed frustration at being asked to provide this type of information. They indicated that many ideas, such as data collection, summarizing data, and the category of "Why Statistics?", should continue throughout the entire course. Two comments worth noting follow:

- "I agree with the comment that filling out this just reconciles me to thinking in the same old probability based way. I am trying hard to get away from that. These categories all run together in a way that is impossible to decode into neat little discrete amounts."
- "I'm trying to cooperate here, but this is really not how I teach! Most of these topics occur throughout the course."

Another panelist cautioned not to take these numbers too seriously. Rather the numbers can serve as an approximate guide.

The recommended time allotment for each topic shown in Table 3 indicated that on average most time should be spend on inferential statistics such as estimation and hypothesis testing. There seemed to be a slight preference for estimation over hypothesis testing, which could reflect the growing support for use of confidence intervals. The topic of probability received less emphasis than techniques for data analysis. These results differ somewhat from other studies, and it is here that the numerical results start to show a slight departure from a traditional course. Recommended course outlines for mathematics majors (Alo, 1983) and engineering majors (Hogg, 1985) both recommended spending longer amounts of time on probability issues. The difference could be due in part to the greater amount of mathematical training by these two groups of students, to the different course prerequisites, and to the fact these reports are somewhat older. There is growing support in the literature and in supplementary course materials to shift the emphasis in introductory statistics courses away from a mathematical/probability based approach and toward a stronger data based approach (Gnanadiskan, Scheaffer, Watkins, & Witmer, 1997; Scheaffer, Gnanadesikan, Watkins, & Witmer, 1996; Witmer, 1992).

Instructional Strategies

The shift in emphasis toward a data driven approach is reflected in the responses concerning instructional approaches. As shown in Table 4 the most favored approach was the so-called data-based approach, followed by the computer-based approach. Depending on how the computer was used, whether for managing large data sets or for generating simulations to illustrate probability concepts, the approach could be used either way. The category of "Other Suggested Approaches" included student laboratories and projects, experimental design, and case studies, which would all place heavy emphasis on data.

As shown in Table 5, the most preferred classroom activity was lecture, followed by discussion. Some panelists indicated that they like to do student projects and presentations but are limited by class size and time. Many said that the type of classroom activity would depend heavily on the size of the class.

Conclusion

The present study was an attempt to address some of the needs of introductory statistics courses by providing a format in which peer recognized leaders in statistics education could share ideas and reach consensus on a number of topics. One particular aspect of a Delphi study is that the validity of the study rests strongly on the formation of the panel of experts used to reach consensus. The method used to form this panel was unique in that, rather than depending on the expertise of one or two people, members of the community of statistics educators were asked to nominate the panel members. The resultant panel was thus composed of peer recognized and peer respected experts in statistics education. All panel members were in statistics or closely related fields, thus showing content knowledge. Over three quarters of the panel members were associated with a college or university, thus associated with the teaching of statistics. The remaining panel members were associated with companies and businesses which are employers of students, thus associated with educational outcomes. The panel members themselves were highly motivated to participate in this study, as indicated by the high response rate on each round of questionnaires. The number of personal comments made on the questionnaires also indicated a high level of interest. One panel member said, "I think this is an interesting exercise and look forward to your making sense of it."

The results of the study could have implications on textbook selection, curriculum, instructional methodology, and evaluation. The list of essential topics, along with the recommended time for each topic, could be an invaluable tool for new and experienced teachers. The suggested instructional approaches showed a strong appreciation for a data-based approach. There are many excellent resources, both in print and on the World Wide Web, to support and enhance such an approach.

Statistical knowledge is important in society today. In recognition of this, many students are required to take at least one statistics course. It is important that the curriculum for the statistics course be relevant, useful, and motivational. The results of the present study should aid in the development of such courses.

References

Allen, R., Efird, J., & Eliasziw, M. (1990). A symposium on statistical education for the future. *American Statistical Association 1990 Proceedings of the Section on Statistical Education* (pp. 84-89). Alexandria, VA: American Statistical Association.

Alo, R. A. (1983). Statistics in the two-year curriculum. In A. Ralston, & G. S. Young (Eds.), *The Future of College Mathematics: Proceedings of a Conference/Workshop on the First Two Years of College Mathematics* (pp. 145-151). New York: Springer-Verlag.

Ames, M. H., Clason, D. L., & Urguhart, N. S. (1990). An historical perspective on introductory statistics texts. *American Statistical Association 1990 Proceedings of the Section on Statistical Education* (pp. 188-193). Alexandria, VA: American Statistical Association.

Anderson, C. W., & Loynes, R. M. (1986). University statistics--what are we trying to teach and how? *Proceedings of the Second International Conference on Teaching Statistics* (pp. 317-321). Victoria, British Columbia, Canada: University of Victoria Conference Services.

Becker, M. J. (1996). A look at the literature (and other resources) on teaching statistics. *Journal of Educational and Behavioral Statistics, 21*(1), 71-90.

Borg, W. R., & Gall, M. D. (1983). *Educational research: An Introduction* (4th ed.). New York: Longman.

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician, 50*(1), 69-78.

Cobb, G. W. (1987). Introductory textbooks: A framework for evaluation. *Journal of the American Statistical Association, 82*, 321-339.

Cobb, G. W. (1993). Reconsidering statistics education: A National Science Foundation conference. *Journal of Statistics Education* [On-line serial], 1(1). Available: http://www.stat.ncsu.edu/info/jse/

Foreman, B. A., Brown, W. E., & Behrens, J. T. (1992, April). *Trends in the organization of introductory statistics textbooks.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Garfield, J. (1995). How students learn statistics. *International Statistical Review, 63*(1), 25-34.

Gaudard, M., & Hahn, G. J. (1991). An undergraduate concentration in applied statistics for mathematics majors. *The American Statistician, 45*(2), 115-120.

Giesbrecht, N. (1996). *Strategies for developing and delivering effective introductory-level statistics and methodology courses.* (ERIC Document Reproduction Service No. ED 393 668)

Giesbrecht, N., Sell, Y., Scialfa, C., Sandals, L., & Ehlers, P. (1994). *A study of the feasibility of an across-university course in statistics and methodology.* Unpublished manuscript. University of Calgary.

Gnanadesikan, M., Scheaffer, R. L., Watkins, A. E., & Witmer, J. R. (1997). An activity-based statistics course. *Journal of Statistics Education* [On-line serial], *5*(2). Available: http://www.stat.ncsu.edu/info/jse/

Grady, C. S., Looney, S. W., & Steiner, R. P. (1994). A study of biostatistics requirements in medical schools in the United States. *American Statistical Association 1994 Proceedings of the Section on Statistical Education* (pp. 263-265). Alexandria, VA: American Statistical Association.

Grosof, M. S., & Sardy, H. (1990). Introductory statistics in small colleges: Diversity or chaos? *American Statistical Association 1990 Proceedings of the Section on Statistical Education* (pp. 84-89). Alexandria, VA: American Statistical Association.

Hogg, R. V. (1985). Statistical education for engineers: An initial task force report. *The American Statistician, 39*(3), 168-175.

Hogg, R. V. (1992). Towards lean and lively courses in statistics. In F. Gordon & S. Gordon (Eds), *Statistics in the twenty-first century* [MAA Notes No. 26] (pp. 3-13). Washington, DC: Mathematical Association of America.

Jackson, T. (1991). [Mathematics topics recommended in a one year college preparatory non-calculus based statistics course and instructional emphasis: Mean percentages of time that should be devoted to instructional methodology]. Unpublished raw data. (Available from Tess Jackson, Winthrop College, Rock Hill, SC 29733)

Linstone, H. A., & Turoff, M. (Eds.). (1979). *The Delphi Method Techniques and Applications.* Reading, MA: Addison-Wesley Publishing Company.

Lopez, S. D., & Mertens, D. M. (1994). Survey of the SIG: Professors of educational research teaching practices 1992-93. In M. E. Ware & C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, M. T., & Johnson, G. E. (1992). *Research priorities for research and development: Articulating areas for increased investigative attention.* (ERIC Document Reproduction Service No. ED 353 410)

Moore, D. S. (1985). *Statistics concepts and controversies* (2nd ed.). New York: W. H. Freeman and Company.

Neter, J. (1989). Undergraduate statistics service courses in the years ahead. *American Statistical Association 1989 Proceedings of the Section on Statistical Education* (pp. 29-31). Alexandria, VA: American Statistical Association.

Oathout, M. J. (1995). *College students' theory of learning introductory statistics: Phase one.* (ERIC Document Reproduction Service No. ED 391 841)

Scheaffer, R. L., Gnanadesikan, M., Watkins, A., & Witmer, J. R. (1996). *Activity-based statistics: Student guide.* New York: Springer-Verlag.

Somers, K., Baker, G., & Isbell, C. (1984, May). How to use the Delphi technique to forecast training needs. *Performance and Instruction Journal*, 26-28.

Vere-Jones, D. (1995). The coming of age of statistical education. *International Statistical Review, 63*(1), 3-23.

Witmer, J. R. (1992). *Data analysis: An introduction.* New Jersey: Prentice Hall.

Appendix

Delphi Panel Members Nominated by National Survey (n=194)

Donald L. Bentley, Professor, Mathematics Department, Pomona College, Claremont, CA.

John Boyer, Statistics Department, Kansas State University, Manhattan, KS.

George W. Cobb, Prof Stat/Dean of Studies, Math/Stats/Comp, Mt Holyoke College, South Hadley, MA.

Jacquelin E. Dietz, Stat Dept, North Carolina State University, Raleigh, NC.

J. Leroy Folks, Statistics Department, Oklahoma State University, Stillwater, OK.

Berton H. Gunter, Principal and Owner, B. H. Gunter & Assoc., Hopewell, NJ.

Donald Guthrie, Psychiatry Dept, University of California, Los Angeles, CA.

David K. Hildebrand, Stat Dept, Univ of Pennsylvania, Philadelphia, PA.

Robert V. Hogg, Professor, Stat & Actuarial Sci Dept University of Iowa, Iowa City, IA.

Carl J. Huberty, Prof Educ Psychol, Univ of Georgia, Athens, GA.

Dallas E. Johnson, Prof & Consult Stat Dept, Kansas State Univ, Manhattan, KS.

James E. Kearis, 6640 South Williams Circle W, Littleton, CO.

James L. Kepner, Prof/Chair Math & Stat Dept, St Cloud St Univ, St. Cloud, MN.

James Landwehr, Stat Models & Methods Research Dept, AT&T Bell Labs, Murray Hill, NJ.

William R. Loeffler, President, The Loeffler Group Inc., Toledo, Ohio.

Edward R. Mansfield, Mgmt Sci & Stat Dept, Univ of Alabama, Tuscaloosa, AL.

Paul Minton, 2626 Stratford Road, Richmond, VA.

David S. Moore, Professor Statistics Department, Purdue University, West Lafayette, IN.

Thomas L. Moore, Assoc Prof, Math & Cmptr Sci Dept, Grinnel College, Grinnel, IA.

Lincoln Moses, Statistics Department, Stanford University, Stanford, CA.

Federick Mosteller, Stat Dept, Harvard Univ Sci Ctr, Cambridge, MA.

Walter R. Pirie, Statistics Department, VA Polytec & State University, Blacksburg, VA.

Richard L. Scheaffer, Statistics Department, University of Florida, Gainsville, FL.

Lawrence A. Sherr, Chancellor Clb Teaching Professor, School of Business, University of Kansas, Lawrence, KS.

John Skillings, Math & Stat Dept, Miami University, Oxford, OH.

Robert W. Stephenson, Stat Dept, Iowa State University, Ames, IA.

Bruce Thompson, Educational Psychology Department, Texas A&M University, College Station, TX.

Jeffrey A. Witmer, Provost Office, Oberlin College, Oberlin, OH.

Douglas A. Zahn, Statistics Department, Florida State University, Tallahassee, FL.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 1, 61-68

Modeling Asymmetric Hypotheses with Log-Linear Techniques

Frank Lawrence, Gerald Halpin, and Glennelle Halpin
Auburn University

This is a didactic article on the use of log-linear modeling techniques. The paper focuses on asymmetric modeling. The technique is illustrated using data obtained from a sample of 868 pre-engineering students who enrolled in a major southeastern university in 1991. The asymmetric modeling method is used to evaluate the associations among measures of ethnicity, gender, and admission status -- areas of interest to the college's administration.

As is generally true with quantitative data, analyzing associations among measures is at the core of a qualitative data study. Most analysis of categorical or count-type data begins with the cross-classified contingency table. The table may be formed using one-way, two-way, three-way, and higher-order associations among the observed variables (Tabachnick & Fidell, 1996). The cross-classified contingency table provides a visual display of the data and a prefatory glimpse of associations being considered.

Simple tables may be analyzed using the Pearson chi-square statistic. For more complicated tables, another analytic technique is required (Norusis, 1994). A technique proven to work well is log-linear modeling (Agresti, 1996; Everitt, 1992; Fienberg, 1980; Fox, 1997). Log-linear models represent an approach especially designed for the examination of categorical data (Haberman, 1978).

Log-linear is a technique for examining associations among different observed characteristics. The purpose of this paper is to review one type of log-linear model, the asymmetric model, and to show its application to education research. To accomplish this objective, we begin by briefly reviewing log-linear models. Following the review of log-linear models, we focus on a form of asymmetric inquiry called logit modeling. Throughout the paper, we make use of data collected at a major southeastern university to illustrate the technique. We conclude by discussing measures of fit and interpretations of this type model.

Log-Linear Models

Log-linear models form a genre of techniques for solving particular research problems. Researchers typically adopt one of two approaches to problem solving. The approaches emanate from the researcher's hypotheses. Hypotheses may be generally classified as either symmetric or asymmetric. A symmetric hypothesis is one that posits the presence of an association between variables. On the other hand, an asymmetric inquiry is one that not only posits an association but also defines the expected direction of association (Kennedy, 1992).

Log-linear modeling is able to address both types of inquiry. In symmetric log-linear modeling the cell frequency is the raison d'être for the model. The researcher seeks to design a model that will reproduce the observed cell frequencies. There is no attempt to label one of the variables as dependent or independent. The objective in the symmetric analysis is simply to identify associations (Norusis, 1994). Models are evaluated according to the dual criteria of ability to reproduce observed frequencies and parsimony (Demaris, 1992; Everitt, 1992; Tabachnick & Fidell, 1996; Wickens, 1989).

Once an acceptable model is identified, it is interpreted in a proportional sense. For example, imagine that an association is detected between people's gender and their ability to rotate an object mentally. For a symmetric inquiry, the conclusion might be that gender implies a differential ability to cerebrate.

Conclusions regarding significant symmetric findings are antiseptic and somewhat unsatisfying. The dissatisfaction likely stems from most researchers being more experienced with analytic methods that result in a definitive statement about the orientation of the relation. Hence, the feeling at the end of a symmetric analysis that the research is not complete.

Asymmetrical Models

Conversely, asymmetric models are designed to address antecedent relations. Thus, the asymmetric model

---

Frank Lawrence is a doctoral candidate and graduate teaching assistant in the Department of Educational Foundations, Leadership, and Technology at Auburn University. Gerald Halpin and Glennelle Halpin are Professors in the Department of Educational Foundations, Leadership, and Technology at Auburn University. Correspondence concerning this paper should be directed to Frank Lawrence, 4080 Haley Center, Auburn University, Auburn, AL 36849-5221 or by e-mail to halpige@mail.auburn.edu.

posits a causal relationship. While the symmetric approach to inquiry is bilateral, the asymmetric approach is distinctly unidirectional. Hence, this mode of inquiry may be used similar to an analysis of variance. Using an asymmetric model, the researcher can determine if subjects differ in their response rates over distinct levels of an independent variable. In fact, in more complicated designs employing multiple explanatory (independent) variables and one response (dependent) variable, researchers can analyze main effects and interactions (Demaris, 1992; Kennedy, 1992).

Asymmetrical models are the more frequently encountered ilk of log-linear models (Kennedy, 1992). Asymmetric models are the models of choice because they are able to satisfy investigator's needs. Most investigators begin their inquiry with a hypothesis. The hypothesis will usually state a relationship. For example, the null hypothesis might be that there is no difference in the rate of admission for people of different ethnic backgrounds or gender. There is a definite response variable and clearly defined explanatory variables. The analyst seeks to determine the validity of the hypothesis through some sort of data analysis. The asymmetric model is particularly well suited to analyzing this type of unidirectional relations.

Asymmetric inquiries make use of a sept of the general log-linear model called the logit model. In the logit model, the dependent variables are called the logit or response variables and the independent variables are called explanatory variables (Wickens, 1989). We adhere to this nomenclature throughout the paper. As Wickens (1989) indicated, the terms independent and dependent imply a cause and effect relationship. Use of independent and dependent variables can be misleading. Because the asymmetric modeling technique does not require an experimental design and data may be obtained without any action on the part of the researcher to control the data generation process, the use of dependent and independent variable would constitute an inapplicable designation.

The name, logit, means the log odds or natural logarithm of the odds ratio (Demaris, 1992; Fox, 1997; Goodman, 1972; Kennedy, 1992). We have included two appendices in this paper to illustrate odds and logs. Appendix A contains a brief synopsis of odds ratios and their calculations; Appendix B explains by example the behavior of natural logarithms. Appendix A uses the data set discussed below to illustrate the calculations.

Logit models. Imagine two variables hypothesized to be associated. One of these variables (B) is designated as the response variable while the other variable (A) is thought to explain changes in B. To capture the dependency of B on A, a simple function is proposed:

$$B_i = \alpha + \beta A_i + \varepsilon_i \qquad (1)$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, error terms are assumed to be independent and A is random and independent of ε.

The expected value of B_i is actually either 0.0 or 1.0. When B_i takes on only one of two values, the linear regression model denoted by Equation 1 is constrained. Figure 1 shows the result of this constraint on the expected value of the response variable.

Figure 1. Scatter plot of dichotomously scored dependent variable.

The difficulty with the model originates from this specification. Because B_i can become only a 0.0 or 1.0, ε_i is dichotomously distributed. Thus, the distribution of ε_i violates one of the assumptions fundamental to multiple linear regression.

Equation 1 can be rewritten as a probability function. Assume the probability that B_i takes on a value of 1.0 is π_i. Then

$$\pi_i = \alpha + \beta A_i \qquad (2)$$

where π_i is defined as follows:

$$\pi_i \equiv \Pr(B_i) \equiv \Pr\langle B = 1 \mid A = a_i \rangle$$

A central difficulty with the linear probability model of Equation 2 is its inability to ensure the response variable stays between 0.0 and 1.0. This shortfall can be corrected by using a positive monotone function to map the linear predictor into a unit interval. A probability distribution function will satisfy this need. Re-writing Equation 2

$$\pi_i = P(\alpha + \beta A_i) \qquad (3)$$

where the probability distribution function is selected in advance and the intercept and slope are estimated.

The transformation of the probability distribution function expressed in Equation 3 is commonly made using the logistic distribution. The result is a linear logit model expressed as follows:

$$\pi_i = \Lambda(\alpha + \beta A_i) = \frac{1}{1 + \exp\left[-\left(\alpha + \beta A_i\right)\right]} \qquad (4)$$

where Λ represents the logistic distribution. The inverse linear transformation of the function, $\Lambda^{-1}(\pi)$, is the log odds ratio. The transformation can be seen more clearly by rewriting Equation 4

$$\frac{\pi_i}{1 - \pi_i} = \exp(\alpha + \beta A_i)$$

The response ratio is the odds that $B_i = 1$. In this form, there is a difficulty with the relationship; that is, there is no upper bound on the odds ratio. As written, the odds ratio's upper bound is infinity. However, taking the natural logarithm of both sides of the equation yields a response variable bounded by 0.0 and 1.0.

$$\ln \frac{\pi_i}{1 - \pi_i} = \alpha + \beta A_i \qquad (5)$$

Equation 5 is an expression of a logit model. It is linear and additive for the log odds ratio. The parameter, β, reveals the slope of the relationship between the log odds of $\pi_i/1-\pi_i$ and A. Thus, a change of one unit in A_i results in a change of β in the log odds. Alternatively, a unit change in A_i multiplies the odds by e^{β}.

In this section we have derived and explained the logit model. This is the model used for asymmetric analysis. The next section describes the data set we will use to illustrate this modeling technique.

Data set. This data set was obtained from the college of engineering at a major southeastern state university. Student retention is a goal at this university. The recent focus on student attrition has caused the administration concern (cf. Halpin, Halpin, Benefield, & Walker, 1997; Tinto, 1993).

A study was commissioned to investigate pre-engineering student retention. The study analyzed students' success in their quest for admission to the college of engineering. At the end of their second year, students in the pre-engineering program were admitted to either the college of engineering or they were not. Students that are not admitted come from one of two categories. The first of these categories is for students not admitted because their grade point average was less than the threshold value for admission of 2.2. The second category is for students who had a grade point average equal to or above 2.2 but elected not to enter the college of engineering. Thus, the outcome is trichotomous making grade point average an excellent candidate for cross-classified contingency table analysis.

The study consists of the 1991 pre-engineering entering class (N=868). Of the 868 pre-engineering students, 667 were classified as male, 201 females. This class of pre-engineering students consisted of 745 (85.8%) people who reported their ethnic background as Caucasian, 98 (11.3%) African American, 15 (1.7%) Asian, 5 (.6%) Hispanic, 3 (.3%) non-resident alien, and 2 (.2%) American Indian. Because few students reported being from an ethnic group other than Caucasian, the categories for ethnicity were collapsed to two: Caucasian and People of Color. Thus, the selected categories for the ethnic variable are mutually exclusive and exhaustive.

The administration at the college of engineering was interested in investigating the relationship between admission status and ethnic background. Admission status was categorized as "admitted," "not admitted with GPA ≥ 2.2," or "not admitted with GPA < 2.2." However, for didactic purposes the table is initially portrayed as a two-by-two cross-classified contingency table comparing two levels of ethnic background (Caucasian and People of Color) with two levels of admission status (admitted or not admitted). This information is shown in Table 1.

Table 1
Cross-Classification of Ethnicity by College of Engineering Admission Status for 1991 Entering Students

		Ethnicity	
Admission status		Caucasian	People of Color
Admitted	Count	404	45
Not Admitted	Count	341	78
Total	Count	745	123

Logit example. For these data, consider that admission status is a response variable (B) while ethnicity (A) is an explanatory variable. Let π denote the probability of admission. The logit model may now be described as the ratio of the odds of being admitted to the odds of not being admitted given ethnic background. The model may be portrayed as follows:

$$\ln\frac{\pi_i}{1-\pi_i}=\alpha+\beta A_i$$

where A_i stands for different levels of ethnicity. Whenever a logit model is fitted to the data, only the parameters that include the response variable are in the model (Norusis, 1994).

The first model that is considered for this analysis is called a saturated model. By saturated, we mean that it contains all possible terms relating the variables. Hence, it is expected to reproduce exactly the observed frequencies. Furthermore, it cannot be tested, because all the degrees of freedom are used to construct the model. For these reasons, saturated models are generally considered uninteresting (Agresti, 1996; Norusis, 1994; Wickens, 1989). Still, saturated models are useful as baselines for comparison of other, more parsimonious models. For this reason, we begin with the analysis of this necessary, but uninspiring, model.

A separate depiction of the logit model makes use of cell frequencies. Using this format, the saturated model for this test is as follows:

$$\ln\left[\frac{F_{11}}{F_{21}}\right]=\lambda^{\mathrm{d}}+\lambda^{\mathrm{C\ d}}$$

where *C* represents Caucasian and *d* indicates admitted to the college of engineering, and F_{ij} the appropriate cell count. This model is fit to the data in the two-by-two cross-classified contingency table shown in Table 1.

As expected, this model fits the data perfectly. The model is designed to yield the natural log odds of being admitted to the college of engineering for different levels of ethnicity. The parameter of interest is λ^{Cd}, which is the term for the interaction between ethnicity and admission. The term shows how admission status varies across levels of ethnicity. This term has a value of .7147. The value represents the change in log odds of admission for a unit change in ethnicity. Hence, it is interpreted as the difference in the log odds between a Caucasian being in the admitted category and a Person of Color being in the same category of the response variable. If the log odds had been 0, then the odds ratio would have been 1.0 indicating that ethnic background had no relationship with admission status. The fact that the log odds is different from 0 indicates there is a relationship between ethnic background and admission to the college of engineering. Exponentiation of the log odds reveals the impact of ethnic background on admission to the college of engineering. In this example, the odds of a Caucasian being admitted are 2.05. The interpretation of the odds indicate that it is 105% more likely that someone admitted to the college of engineering is Caucasian than a Person of Color.

The two-by-two contingency table does not allow for investigation of more complex models because of its limited dimensions. In the two-by-two table, all variables in the table are required to construct the most basic model. If the table is expanded to a two-by-two-by-two design, models that are more elaborate may be proposed and tested.

Table 2 illustrates the more complex design just mentioned. It is a cross-classified contingency table showing the relationship among gender, ethnicity, and admission status.

Table 2
Cross-Classification of Gender by Ethnicity by College of Engineering Admission Status for 1991 Entering Students

Ethnicity	Gender	Admitted	Not Admitted	Total
Caucasian	Female	73	76	149
	Male	331	265	596
People of Color	Female	20	32	52
	Male	25	46	71

The information in Table 2 is used to test the hypothesis that Caucasian females are more likely to be admitted to the college of engineering. The saturated model for this test is as follows:

$$\ln\left[\frac{F_{111}}{F_{211}}\right]=\lambda^{\mathrm{d}}+\lambda^{\mathrm{C\ d}}+\lambda^{\mathrm{F\ d}}+\lambda^{\mathrm{C\ F\ d}}$$

where the exponent, *F*, represents female, *C* Caucasian, *d* admitted to the college of engineering, and F_{ijk} the appropriate cell count. As expected, this model fits the data perfectly.

The analysis renders parameter estimates for each term in the model. These estimates are most useful in determining the direction of difference as well as the magnitude. Yet, by themselves, they do not tell us if the suspected differences are statistically significant. To establish significance, a *z*-test is employed. The *z*-test of interest for this hypothesis is the test that the parameter is zero.

The *z*-test is of the form

$$z(\lambda_{ijk}^{\mathrm{CFd}}) = \frac{\hat{\lambda}_{ijk}^{\mathrm{CFd}}}{ASE(\hat{\lambda}_{ijk}^{\mathrm{CFd}})}$$

where *ASE* is the asymptotic standard error (Agresti, 1996). For this analysis, the *z*-test indicates the parameter is not significant ($z = 1.00$). Thus, we fail to reject the null hypothesis that Caucasian females are not advantaged in admission status.

Although this parameter estimate is not significant, for pedagogical purposes we continue the analysis. Using the parameter estimate and the *ASE*, it is possible to construct confidence intervals around the parameter estimate. The 95% confidence interval around this parameter estimate is log odds –1.06 to .83. The confidence interval is calculated using the following formula:

$$\hat{\lambda}_{ijk}^{\mathrm{CFd}} \pm z_{\alpha/2} ASE * \hat{\lambda}_{ijk}^{\mathrm{CFd}}$$

Given that the interaction term is not significant, the other terms in the model are evaluated to ascertain if a less complex model might prove sufficient to represent the observed frequencies. The evaluation indicates that the ethnicity by admission status term is statistically significant ($z = 2.99$). No other two-way or more complex interaction terms are significant.

Measures of Model Significance

To determine if the logit model is a viable representation of the observed data, two measures of significance are referenced. The first of these is the Pearson chi-square statistic; the second is the Fisher likelihood ratio chi-square statistic. Both statistics evaluate model fit. That is, they examine the difference between the estimated values and the observed values to determine if the discrepancy is significantly different from zero.

The formula for the Pearson chi-square is

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O-E)^2}{E} \tag{8}$$

where O is the observed count in the cell in the i^{th} row and j^{th} column and E is the expected count. The chi-square test evaluates the null hypothesis that there is no difference between the observed and expected frequencies in each cell.

The formula for the Fisher likelihood ratio chi-square statistic (G^2) is

$$G^2 = 2\sum o_{ij} (\ln \frac{o_{ij}}{E_{ij}}) \tag{9}$$

where *ln* stands for the natural logarithm. Like Pearson's chi-square, the likelihood ratio chi-square is chi-square distributed. Both the Pearson chi-square and the likelihood ratio chi-square are asymptotically equivalent. Moreover, the likelihood ratio chi-square statistic has the following property:

$$G_T^2 = G_i^2 + G_j^2 + G_{ij}^2 \tag{10}$$

The formula signifies that the total chi-square value is the sum of the chi-square for the main row and column effects as well as the interaction between the column and row variables.

Table 3 shows the chi-square values for the logit model with only the gender and ethnicity main effects.

Table 3
Goodness of Fit Statistics for Ethnicity by Admission Status Logit Model

Type	Chi-square	DF	Sig.
Likelihood Ratio	2.6013	2	.2723
Pearson	2.6137	2	.2707

There is a similarity between the Pearson chi-square value and the likelihood ratio chi-square value. This similarity exists because there is no overlap or shared variance to account for in the model.

Neither the Pearson chi-square nor the likelihood ratio chi-square are significant at the alpha=.05 level. The interpretation is that the model provides an adequate fit to the data.

There are two general classifications of chi-square values generated in log-linear analysis. These classifications are the residual chi-square and the component chi-square. The residual chi-square is the chi-square value resulting from an evaluation of model fit. The residual chi-square tests the null hypothesis that the residual values resulting from fitting a model are zeros in the population. The component chi-square is the difference in residual chi-square values between two nested models. The degrees of freedom for the component chi-square are the difference in the degrees of freedom for the models being evaluated.

To identify the model that best satisfies the criteria of parsimony and fit, the models are arranged in hierarchical order with their associated residual chi-square values. Beginning with the saturated model, terms are eliminated until a significant component chi-square value is obtained. At this juncture, the elimination is terminated and the model is interpreted. The result of applying this process to the data in Table 2 is shown in Appendix B, Table 4.

Discussion

A reasonable logit model is one that serves to explain differences in cross-classified contingency table cell counts. In other words the model fits the observed data as indicated by the measures of fit and is defensible. In the example just discussed, the model was reasonable because it addressed a specific research question and the interpretation of the results was sound.

By choosing to use logit models, the researcher is inferring a concern with response patterns. Sampling can distort the interpretation of response patterns. For example, if the proportion of females sampled is twice that of males, it is difficult to ascertain by observation if frequency differences are due to gender differences or sampling vagaries. Indeed, the marginal distribution of subjects is irrelevant to the problem facing the researcher. Therefore, they should not be permitted to influence the analysis.

Differences in the main marginal effects should be acknowledged and controlled during the investigation. Control can be accomplished by incorporating the effects into the model. In the example, a term representing gender was included to ensure that differences in sample patterns did not sway the outcome. If the model with the gender-by-status term had been selected as the best representation of the cross-classified contingency table data, then our conclusion might have been that females are twice as likely to be admitted to the college of engineering as males when controlling for ethnicity. Nevertheless, this term was not significant; therefore, this interpretation was not advanced.

To be a legitimate model, the logit must contain all terms that reflect potential differences not related to proportion of response over the logit variable. When there are different proportional response rates over the logit variable, terms reflecting these differences should be included in the model. Otherwise, the model is flawed and any conclusions drawn from the analysis will be suspect.

In sum, reasonable logit models contain terms representing the main marginal effects as well as higher order interactions when those interactions include theoretically based explanatory variables. In the model used for Table 2 data, terms were incorporated to represent gender, ethnicity, and the interaction effect of ethnicity upon admission status. The last term was included because it was the focus of the inquiry and was theoretically grounded.

To locate reasonable logit models, a hierarchy of models may be proposed. The logit conversion can be extended to the entire family of nested models. Symbolically, the hierarchy of models representing the information in Table 2 would be as follows:

$$\text{logit} = \hat{\lambda}_i^{d}$$

$$\text{logit} = \hat{\lambda}_i^{d} + \hat{\lambda}_{ij}^{Cd}$$

$$\text{logit} = \hat{\lambda}_i^{d} + \hat{\lambda}_{ij}^{Cd} + \hat{\lambda}_{ik}^{Fd}$$

$$\text{logit} = \hat{\lambda}_i^{d} + \hat{\lambda}_{ij}^{Cd} + \hat{\lambda}_{ik}^{Fd} + \hat{\lambda}_{ijk}^{CFd}$$

where the subscript denotes the level and the superscript denotes the variables with labels as indicated above.

In a model hierarchy such as depicted here, each preceding model is nested within the succeeding one. For example, the model with the main effects for ethnicity ($\hat{\lambda}_{ij}^{Cd}$) and gender ($\hat{\lambda}_{ik}^{Fd}$) is nested within the saturated model, the one containing the interaction effect ($\hat{\lambda}_{ijk}^{CFd}$). Likewise, the model containing only the main effect for the ethnicity variable is nested within the model containing main effects for ethnicity and gender. Finally, the model containing one parameter ($\hat{\lambda}_i^{d}$) is nested within the model having the ethnicity-by-admission status parameter. This hierarchical arrangement allows for computation and interpretation of the most parsimonious model.

Returning to the data in Table 2, all theoretically important nested models should be examined to determine which ones provide valuable and viable information. Models should be theoretically significant as well as statistically significant. It is extremely difficult to interpret models that show statistical significance but have no theoretical foundation. Therefore, nested models were tested that included main effects for gender and ethnicity as well as interaction effects between all explanatory variables. Only the effect of ethnicity upon admission status was statistically significant ($z = 2.99$). Hence, the odds of a Caucasian being admitted to the college of engineering in 1991 were significantly better than for a Person of Color. The significance of the effect means that there were factors, other than chance, causing the difference in odds of admission. Further investigation is required to determine the nature of those factors.

Conclusion

Asymmetric analysis is a valuable addition to the researcher's arsenal. The asymmetric inquiry has three distinct advantages when applied to qualitative data analysis. First, it exhibits power paralleling that of analysis of variance inquiries (Agresti, 1996). Second, the results of an asymmetric inquiry are generally interpreted in a fashion closely aligned with an analysis of variance, which makes the statistics garnered from applying this tool easier to interpret. Finally, the asymmetric approach to examination of unproved theory is one akin to classical hypothesis testing. Therefore, it may be more comfortable for many researchers.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Demaris, A. (1992). *Logit modeling: Practical applications*. Newbury Park: Sage.

Everitt, B. S. (1992). *The analysis of cross-classified contingency tables* (2nd ed.). London: Chapman & Hall.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: The MIT Press.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks: Sage.

Goodman, L. A. (1972). A modified multiple regression approach to the analysis of dichotomous variables. *American Sociological Review, 37*, 28-46.

Gravetter, F. J., & Wallnau, L. B. (1995). *Essentials of statistics for the behavioral sciences* (2nd ed.). Minneapolis: West Publishing Company.

Haberman, S. J. (1978). *Analysis of qualitative data.* Vol. 1. New York: Academic Press.

Halpin, G., Halpin, G., Benefield, L. D., & Walker, W. F. (1997, March). *Retention in engineering education: Longitudinal race and gender differences.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Kennedy, J. J. (1992). *Analyzing qualitative da5ta: Log-linear analysis for behavioral research* (2nd ed.). New York: Praeger.

Norusis, M. J. (1994). *SPSS advanced statistics 6.1.* Chicago: SPSS.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). Fort Worth: Harcourt Brace.

Ramanathan, R. (1989). *Introductory econometrics with applications.* San Diego: Harcourt Brace Jovanovich.

Scheaffer, R. L., Mendenhall, W., & Ott, L. *Elementary survey sampling* (3rd ed.). Boston: Duxbury Press.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). California State University, Northridge: Harper Collins.

Tinto, V. (1993). *Leaving college: Rethinking the causes and curse of student attrition* (2nd ed.). Chicago: University of Chicago Press.

Wickens, T. D. (1989). *Multiway cross-classified contingency tables analysis for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Appendix A

Odds are the relative probability of an event. To illustrate, consider the odds of being admitted to the college of engineering using the data in Table 1. The odds of being admitted are 1.07. The odds ratio is interpreted as a randomly drawn subject from this study who is 7% more likely to be one that is admitted to the college of engineering than one that is not.

Similarly, the odds of being in a particular racial group may be calculated. For example, the odds of being Caucasian in this study are 6.06. Put another way, a randomly selected subject is over six times more likely to be Caucasian than from some other racial group.

To calculate the odds, begin with the number of subjects reported in the category of interest; say admitted. Divide the number admitted by the sum of those assigned to other categories.

When the odds information is limited to one category of a variable, it is said to be conditional. For example, the conditional odds of a Caucasian being admitted to the college of engineering are 1.18. That is, given the subject is Caucasian, she or he has an 18% better chance of being admitted to the engineering program than not being admitted. The odds are conditioned on a characteristic of the subject. In this situation, the odds are conditioned on the subject being Caucasian.

On occasion, the analyst will want to make a comparative statement regarding two categories in the cross-classified contingency table. For example, suppose the analyst wants to compare the odds of a Caucasian being admitted to the engineering program to the odds of a person from another ethnic background being admitted. In this situation, the odds is the ratio of two conditional odds; the odds of the person being admitted given Caucasian to the odds of the person being admitted given some other ethnic background. The odds of admission given that the person is a Caucasian are 1.18, and the odds of admission given another ethnic background are .57. Thus, the odds that the person admitted is Caucasian are 2.05. The interpretation of these odds is that it is 105% more likely the person admitted to the engineering college is Caucasian than a Person of Color.

Appendix B

The purpose of this appendix is simply to illustrate the behavior of natural logarithms. It is included primarily for those readers not familiar with the term or the behavior of the natural logarithms.

Natural logarithms provide a convenient way to transform odds ratios into linear models. For this reason, we illustrate the behavior of the natural logarithm using some of the information contained in the paper.

The natural logarithm of a number is easily obtained using statistical software or a suitable calculator. The natural logarithm, sometimes called the log to the base e, is a transformation process. It allows investigators to change probability models into linear models thus making data manipulation easier and more comfortable.

The table below shows some transformations using the natural logarithm.

Table 4
Table of Natural Logarithms

Base 10	Natural Log
0	undefined
0.1	-2.302585093
0.2	-1.609437912
0.3	-1.203972804
0.4	-0.916290732
0.5	-0.693147181
0.6	-0.510825624
0.7	-0.356674944
0.8	-0.223143551
0.9	-0.105360516
1	0
10	2.302585093
100	4.605170186

Of particular note is the natural logarithm for 0, 10, and 100. The natural logarithm begins undefined for 0. It then starts very negative for values less than 1.0 becoming less negative as it approaches 1.0. After passing 1.0, the natural logarithm becomes positive.

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◆◆

Please enter a subscription to RESEARCH IN THE SCHOOLS for:

Name: ____

Institution: ____

Address: ____

____

			COST
Individual Subscription ($25 per year)	Number of years	_____	_______
Institutional Subscription ($30 per year)	Number of years	_____	_______
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	_____	_______
Back issues for Volumes 1, 2, 3, and 4 ($30 per Volume)	Number of Volumes	_____	_______
TOTAL COST:			_______

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
RESEARCH IN THE SCHOOLS
University of Alabama at Birmingham
School of Education, 233 Educ. Bldg.
901 13th Street, South
Birmingham, AL 35294-1250

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education; many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

* * *

MSERA Membership/Renewal Form

(Please print or type)

Name ____________________

Organization ____________________

Address ____________________

____

Telephone Work: ____________________

Home: ____________________

Fax: ____________________

e-mail: ____________________

Amount Enclosed:	MSERA 1998 Membership ($25 professional, $15 student)	$______
	MSER Foundation Contribution	$______
	TOTAL	$______

Make check out to MSERA and mail to:

Dr. Clifford Hofwolt
MSERA Secretary-Treasurer
Vanderbilt University
Box 330, Peabody College
Nashville, TN 37203

RESEARCH IN THE SCHOOLS
Mid-South Educational Research Association
and the University of Alabama at Birmingham
901 South 13th Street, Room 233
Birmingham, AL 35294-1250

BULK RATE
U.S. POSTAGE
PAID
PERMIT NO. 1256
BIRMINGHAM, AL

MSERA

RESEARCH IN THE SCHOOLS

A nationally refereed journal sponsored by the
Mid-South Educational Research Association
and the University of Alabama at Birmingham.

Volume 5, Number 2 **Fall 1998**

SPECIAL ISSUE
STATISTICAL SIGNIFICANCE TESTING

Introduction to the Special Issue on Statistical Significance Testing 1
Alan S. Kaufman

The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing 3
Thomas W. Nix and J. Jackson Barnette

The Role of Statistical Significance Testing in Educational Research 15
James E. McLean and James M. Ernest

Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals 23
Larry G. Daniel

Statistical Significance and Effect Size Reporting: Portrait of a Possible Future 33
Bruce Thompson

Comments on the Statistical Significance Testing Articles 39
Thomas R. Knapp

What If There Were No More Bickering About Statistical Significance Tests? 43
Joel R. Levin

A Review of Hypothesis Testing Revisited: Rejoinder to Thompson, Knapp, and Levin 55
Thomas W. Nix and J. Jackson Barnette

Fight the Good Fight: A Response to Thompson, Knapp, and Levin 59
James M. Ernest and James E. McLean

The Statistical Significance Controversy Is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin 63
Larry G. Daniel

Title Index, Volumes 1 - 5 67

Author Index, Volumes 1 - 5 71

James E. McLean and Alan S. Kaufman, Editors

RESEARCH IN THE SCHOOLS

Information for Authors

Statement of Purpose

RESEARCH IN THE SCHOOLS (ISSN 1085-5300) publishes original contributions in the following areas: 1) *Research in Practice*--empirical studies focusing on the results of applied educational research including cross-cultural studies, 2) *Topical Articles*--scholarly reviews of research, perspectives on the use of research findings, theoretical articles, and related articles, 3) *Methods and Techniques*--descriptions of technology applications in the classroom, descriptions of innovative teaching strategies in research/measurement/statistics, evaluations of teaching methods, and similar articles of interest to instructors of research-oriented courses, 4) *Assessment*--empirical studies of norm-referenced, criterion-referenced, and informal tests in the areas of cognitive ability, academic achievement, personality, vocational interests, neuropsychological functioning, and the like, and 5) *Other* topics of interest to educational researchers. *RESEARCH IN THE SCHOOLS* is devoted to research conducted in *any* educational setting from a conventional elementary school or high school to a training program conducted within an industry. Likewise, there are no age restrictions on the sample, since the educational settings may include preschools, continuing education classes for adults, or adaptive skills courses in nursing homes. Studies conducted in settings such as clinics, hospitals, or prisons are ordinarily inappropriate for *RESEARCH IN THE SCHOOLS* unless they involve an educational program within such a setting. One goal of *RESEARCH IN THE SCHOOLS* is to provide a training ground for graduate students to learn effective reviewing techniques. Consequently, the journal utilizes a Graduate Student Editorial Board composed mostly of students in educational psychology and educational research. Members of this Editorial Board, each sponsored by a professor, provide supplementary reviews for a selection of submitted articles, and receive both direct and indirect feedback of the quality of these reviews.

Preparing Manuscripts

Authors should prepare manuscripts in accordance with the stylistic rules and guidelines delineated in the *Publications Manual of the American Psychological Association* (4th ed., 1994), which is available from: Order Department, American Psychological Association, PO Box 2710, Hyattsville, MD 20784. Number the pages consecutively. All manuscripts will be subject to editing for sexist language.

Author Identification

Authors should put the complete title of the article on the first text page, but they should exclude their names. Subsequent pages should include only a running head. They should prepare a separate sheet with the complete title of the article and their names and affiliations; this procedure will ensure anonymity in the review process. Authors should supply addresses and phone numbers, and electronic mail addresses and fax numbers (if available), for potential use by the editorial staff and, later, by the production staff. Unless otherwise stated, the first-named author will be sent correspondence, galley proofs, copyright forms, and so forth.

Submission of Manuscripts

Submit manuscripts in triplicate to **James E. McLean, Co-Editor, RESEARCH IN THE SCHOOLS, School of Education, 233 Educ. Bldg., The University of Alabama at Birmingham, 901 13th Street, South, Birmingham, AL 35294-1250**. **Please direct questions to jmclean@uab.edu**. All copies should be clear and readable; dot matrix is acceptable only if it meets these qualities of legibility. Length of the manuscripts, including references and tables, should ordinarily range from about 10 to 40 typed, double-spaced, 8-1/2 X 11-inch pages, using 11-12 point type. Abstracts are limited to 125 words. Brief reports of research are not encouraged. Authors are encouraged to keep a hard copy of the manuscript to guard against loss. It is assumed that all manuscripts submitted for publication are original material and have not been simultaneously submitted for publication elsewhere. When manuscripts are accepted for publication, authors are encouraged to submit the final version on a computer disk along with the hard copy.

Copyright and Permissions

Authors are granted permission to reproduce their own articles for personal use. Others must request permission to reproduce tables, figures, or more than 500 words of text from the editors. Copyright © 1998 by the Mid-South Educational Research Association.

RESEARCH IN THE SCHOOLS

Volume 5, Number 2 **Fall 1998**

EDITORS
James E. McLean, *University of Alabama at Birmingham*
Alan S. Kaufman, *Yale University, School of Medicine*

PRODUCTION EDITOR
Margaret L. Rice, *The University of Alabama*

EDITORIAL ASSISTANT
Michele G. Jarrell, *The University of Alabama*

EDITORIAL BOARD
Gypsy A. Abbott, *University of Alabama at Birmingham*
Charles M. Achilles, *Eastern Michigan University*
J. Jackson Barnette, *The University of Iowa*
Mark Baron, *University of South Dakota*
Robin A. Cook, *Wichita State University*
Larry G. Daniel, *The University of North Texas*
Donald F. DeMoulin, *University of Tennessee–Martin*
Daniel Fasko, Jr., *Morehead State University*
Tracy Goodson-Espy, *University of North Alabama*
Glennelle Halpin, *Auburn University*
Toshinori Ishikuma, *Tsukuba University (Japan)*
JinGyu Kim, *Seoul National University of Education (Korea)*
Jwa K. Kim, *Middle Tennessee State University*
Robert E. Lockwood, *Alabama State Department of Education*
Robert Marsh, *Chattanooga State Technical Community College*
Jerry G. Mathews, *Auburn University*
Charles L. McLafferty, *University of Alabama at Birmingham*
Peter C. Melchers, *University of Cologne (Germany)*
Claire Meljac, *Unité de Psychopathologie de l'Adolescent (France)*
Soo-Back Moon, *Catholic University of Hyosung (Korea)*
Arnold J. Moore, *Mississippi State University*
David T. Morse, *Mississippi State University*
Jack A. Naglieri, *The Ohio State University*
Sadegh Nashat, *Unité de Psychopathologie de l'Adolescent (France)*
Anthony J. Onwuegbuzie, *Valdosta State University*
William Watson Purkey, *The University of North Carolina at Greensboro*
Cecil R. Reynolds, *Texas A & M University*
Janet C. Richards, *The University of Southern Mississippi*
Michael D. Richardson, *Georgia Southern University*
John R. Slate, *Valdosta State University*
Scott W. Snyder, *University of Alabama at Birmingham*
Bruce Thompson, *Texas A & M University*

GRADUATE STUDENT EDITORIAL BOARD
Margery E. Arnold, *Texas A & M University*
Vicki Benson, *The University of Alabama*
Alan Brue, *University of Florida*
Brenda C. Carter, *Mississippi State University*
Jason C. Cole, *California School of Professional Psychology*
James Ernest, *University of Alabama at Birmingham*
Harrison D. Kane, *University of Florida*
James C. Kaufman, *Yale University*
Kevin M. Kieffer, *Texas A & M University*
Pamela A. Taylor, *Mississippi State University*

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 1

Introduction to the Special Issue on Statistical Significance Testing

Alan S. Kaufman
Co-Editor, RESEARCH IN THE SCHOOLS
Clinical Professor of Psychology
Yale University, School of Medicine

The controversy about the use or misuse of statistical significance testing that has been evident in the literature for the past 10 years has become the major methodological issue of our generation. In addition to many articles and at least one book that have been written about the subject, several journals have devoted special issues to dealing with the issues surrounding its use. Because this issue has become so prevalent and it impacts on research in the schools in general and articles published in the *RESEARCH IN THE SCHOOLS* journal as well, James McLean and I--as co-editors of the journal--felt that a special issue that explored all sides of the controversy was in order. To me, personally, the topic is an exciting one. I have published a great many research articles during the past three decades, and often have felt that statistical significance was an imperfect tool. Why should a trivial difference in mean scores or a correlation that begins with a zero be significant simply because the sample is large? Yet, until I began reading articles that challenged the holiness of the birthright of statistical significance testing, I must confess that it never occurred to me to even ask questions such as, "Is there a better way to evaluate research hypotheses?" or "Is statistical significance testing essential to include in a research article?"

This special issue begins with three articles that explore the controversy from several perspectives (Nix and Barnette, McLean and Ernest, and Daniel). These three articles were submitted independently of each other, coincidentally at about the same time, and were peer-reviewed by our usual review process. I then asked the three sets of authors if they would be willing to have their articles serve as the stimuli for a special issue on the topic, and all readily agreed. I then solicited three respondents to the three articles (Thompson, Knapp, and Levin), researchers who seemed to represent the whole gamut of opinions on the topic of the use and possible misuse of statistical significance testing. I asked Bruce Thompson to respond to the articles, even though he had already served as a peer reviewer of these manuscripts, because of his eminence in the field. The three responses to the manuscript follow the three main articles. The special issue concludes with rejoinders from the three initial sets of authors. I believe that you will find the disagreements, none of which are vitriolic or personal, to be provocative and fascinating. Because co-editor James McLean was an author of one of the significance testing articles, he did not participate in editorial decisions with respect to this issue of the journal.

Both Jim McLean and I are very interested in your--the reader's--response to this special issue. We would like to know where our readership stands on the controversial topics debated in the pages of this special issue. We would like to invite you to send us your opinions on the use and misuse of statistical significance testing--what points you agree with and which ones you find not to be very persuasive. We intend to develop a unified policy on this topic for *RESEARCH IN THE SCHOOLS*, which we will base not only on the content of this special issue of the journal, but also on your opinions. We will print every letter that we receive on the topic in the same future issue of our journal that includes our policy statement.

Finally, this issue represents the completion of five years of publication of *RESEARCH IN THE SCHOOLS*. Both author and title indexes are included in this issue to commemorate that accomplishment and make past articles more accessible. In addition, the ERIC Clearinghouse on Assessment and Evaluation catalogs each issue, making *RESEARCH IN THE SCHOOLS* searchable through the ERIC database.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 3-14

The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing

Thomas W. Nix
University of Alabama

J. Jackson Barnette
University of Iowa

Null Hypothesis Significance Testing (NHST) is reviewed in a historical context. The most vocal criticisms of NHST that have appeared in the literature over the past 50 years are outlined. The authors conclude, based on the criticism of NHST and the alternative methods that have been proposed, that viable alternatives to NHST are currently available. The use of effect magnitude measures with surrounding confidence intervals and indications of the reliability of the study are recommended for individual research studies. Advances in the use of meta-analytic techniques provide us with opportunities to advance cumulative knowledge, and all research should be aimed at this goal. The authors provide discussions and references to more information on effect magnitude measures, replication techniques and meta-analytic techniques. A brief situational assessment of the research landscape and strategies for change are offered.

It is generally accepted that the purpose of scientific inquiry is to advance the knowledge base of humankind by seeking evidence of a phenomena via valid experiments. In the educational arena, the confirmation of a phenomena should give teachers confidence in their methods and policy makers confidence that their policies will lead to better education for children and adults. We approach the analysis of experimentation with the tools of statistics, more specifically, descriptive and inferential statistics. Little controversy surrounds the use of descriptive statistics to mirror the various states of nature, however the use of inferential statistics has a long and storied history. Today, there are at least four different schools of thought on inferential significance testing. They are the Fisherian approach, the Neyman-Pearson school, Bayesian Inference, and Likelihood Inference. A full description of each is beyond the scope of this paper, but a complete evaluation of each has been detailed by Oakes (1986). It is fair to state that not one of these inferential statistical methods is without controversy.

We first review the two most popular inferential approaches, the Fisherian and Neyman-Pearson schools, or what has come to be called null hypothesis significance testing (NHST). We then outline some of the major points found in critiques of NHST. Thirdly, we review the changing face of social science research with short primers on effect magnitude measures, meta-analytic methods, and replication techniques. Next, we assess how the development of these methods is coming face-to-face with the shortcomings of NHST. We outline how the primary researcher working on a single study of a phenomena can report more informative information using the same data now used for NHST and at the same time provide his/her study as the raw material for secondary research to be used by a meta-analytic researcher. We conclude with an assessment of the current situation and how change could be facilitated. Through this interchange of ideas and analysis, we can bring some order to what appears to be a chaotic world where the advancement of cumulative knowledge is slowed by a lack of information provided by NHST, misunderstandings about the meaning of NHST results, frustration with conflicting results, and bias in publication policies. Signals in the environment seem to indicate that discussions regarding whether NHST should be banned or not no longer seem to be germane. Rather, the informed stakeholders in the social sciences seem to be abandoning NHST, and with some guidance, we believe the transition to more enlightened statistical methods could be accomplished with minimal disruption.

Development of Null Hypothesis Significance Testing

To better understand how NHST achieved its status in the social sciences, we review its development. Most who read recent textbooks devoted to statistical methods are inclined to believe statistical significance testing is a

---

Thomas W. Nix is a doctoral candidate at the University of Alabama. J. Jackson Barnette is associate professor of Preventive Medicine, Divisions of Community Health and Biostatistics, College of Medicine, University of Iowa. Correspondence regarding this article should be addressed to Thomas W. Nix, 700 Whippoorwill Drive, Birmingham, AL 35244 or by e-mail to tnix@bamaed.ua.edu.

unified, non-controversial theory whereby we seek to reject the null hypothesis in order to provide evidence of the viability of the alternative hypothesis. A *p*-value and an alpha level (α) are provided to determine the probability of the evidence being due to chance or sampling error. We also accept the fact there are at least two types of errors that can be committed in this process. If we reject the null hypothesis, a type I error, or a false positive result, can occur, and if we do not reject the null hypothesis, a type II error, or a false negative result, can occur. Most texts imply NHST is a unified theory that is primarily the work of Sir Ronald Fisher and that it has been thoroughly tested and is above reproach (Huberty, 1993). Nothing could be further from the truth.

The theory of hypothesis testing is not a unified theory at all. Fisher proposed the testing of a single binary null hypothesis using the *p*-value as the strength of the statistic. He did not develop or support the alternative hypotheses, type I and type II errors in significance testing, or the concept of statistical power. Jerzy Neyman, a Polish statistician, and Egon Pearson, son of Karl Pearson, were the originators of these concepts. In contrast to Fisher's notion of NHST, Pearson and Neyman viewed significance testing as a method of selecting a hypothesis from a slate of candidate hypotheses, rather than testing of a single hypothesis.

Far from being in agreement with the theories of Neyman and Pearson, Fisher was harshly critical of their work. Although Fisher had many concerns about the work of Neyman and Pearson, a major concern centered around the way Neyman and Pearson used manufacturing acceptance decisions to describe what they saw as an extension of Fisher's theory. Fisher was adamant that hypothesis testing did not involve final and irrevocable decisions, as implied by the examples of Neyman and Pearson. However, his criticism was not always sparked by constructive scientific debate. Earlier in Fisher's career, he bitterly feuded with Karl Pearson while Pearson was the editor of the prestigious journal, *Biometrika* (Cohen, 1990). In fact, the rift became so great, Pearson refused to publish Fisher's articles in *Biometrika*. Although Neyman and the younger Pearson attempted to collaborate with Fisher after the elder Pearson retired, the acrimony continued from the 1930's until Fisher's death in July, 1962 (Mulaik, Raju, & Harshman, 1997).

Huberty's (1993) review of textbooks outlines the evolution of these two schools of thought and how they came to be perceived as a unified theory. He found that in the 1930s, writers of statistics textbooks began to refer to Fisher's methods, while a 1940 textbook was the first book in which the two types of error are identified and discussed. It was not until 1949 that specific references to Neyman and Pearson contributions were listed in textbooks, in spite of the fact that Neyman and Pearson's work was contemporary to that of Fisher. By 1950, the two separate theories began to be unified in textbooks but without the consent or agreement of any of the originators. By the 1960's the unified theory was accepted in a number of disciplines including economics, education, marketing, medicine, occupational therapy, psychology, social research, and sociology. At the end of the 1980s, NHST, in its unified form, had become so ubiquitous that over 90% of articles in major psychology journals justified conclusions from data analysis with NHST (Loftus, 1991).

Objections to Null Hypothesis Statistical Testing (NHST)

Criticism of NHST provides much evidence that it is flawed and misunderstood by the many who routinely use it. It has even been suggested that dependence on NHST has retarded the advancement of scientific knowledge (Schmidt, 1996b). Objections to NHST began in earnest in the early 1950s as NHST was gaining acceptance. While reviewing the accomplishments in statistics in 1953, Jones (1955) said, "Current statistical literature attests to increasing awareness that the usefulness of conventional hypothesis testing methods is severely limited" (p. 406). By 1970, an entire book was devoted to criticism of NHST in wide ranging fields such as medicine, sociology, psychology, and philosophy (Morrison & Henkel, 1970). Others, including Rozeboom (1960), Cohen (1962), Bakan (1966), Meehl (1978), Carver (1978), Oakes (1986), Cohen (1994), Thompson (1995, November) and Schmidt (1996a), have provided compelling evidence that NHST has serious limiting flaws that many educators and researchers are either unaware of or have chosen to ignore. Below, we examine some of the often quoted arguments. They relate to: a) the meaning of the null hypothesis, b) the concept of statistical power, c) sample size dependence, and d) misuse of NHST information.

The Concept of a Null Hypothesis

In traditional NHST, we seek to reject the null hypothesis (H_o) in order to gain evidence of an alternative or research hypothesis (H_a). The null hypothesis has been referred to as the hypothesis of no relationship or no difference (Hinkle, Wiersma, & Jurs, 1994). It has been argued that, only in the most rare of instances, can we fail to reject the hypothesis of no difference (Cohen, 1988; Meehl, 1967, 1978). This statement has merit when we consider that errors can be due to treatment differences, measurement error and sampling error. Intuitively, we know that in nature it is extremely rare to find two identical cases of anything. The test of differences in NHST posits an almost impossible situation where the null hypothesis differences will be exactly zero. Cohen points out the absurdity of this notion when he states, "... things get downright ridiculous when . . . (the null hypothesis). . . (states) that the effect size is 0, that the proportion of males is .5, that the rater's reliability is 0" (Cohen, 1994). Others have pointed out, "A glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature" (Bakan, 1966). Yet we know that there are tests where the null hypothesis is not rejected. How can this happen given the situation described above? To understand this we turn to the problems associated with statistical power, type I errors, and type II errors in NHST.

Type I Errors, Type II Errors, and Statistical Power

Neyman and Pearson provided us with the two types of errors that occur in NHST. They are type I errors or errors that occur when we indicate the treatment was effective when it was not (a false positive) and type II errors or errors that occur when we indicate there was no treatment effect when in fact there was (a false negative). The probability of a type I error is the level of significance or alpha (α). That is, if we choose a .05 level of significance, the probability of a type I error is .05. The lower the value we place on alpha, for example .01, the more exact the standard for acceptance of the null hypothesis and the lower the probability of a type I error. However, all things being equal, the lower the probability of a type I error, the lower the power of the test.

Power is the probability that a statistical test will find statistical significance (Rossi, 1997, p. 177). As such, moderate power of .5 indicates one would have only a 50% chance of obtaining a significant result. The complement of power (1 - power), or beta (β), is the type II error rate in NHST. Cohen (1988, p. 5) pointed out the weighting procedure the researcher must consider prior to a null hypothesis test. For example, if alpha is set at .001, the risk of a type I error is minuscule, but the researcher may reduce the power of the test to .10, thereby setting the risk of a type II error at (1 - .10) or .90! A power level of .10, as in the previous example, would mean the researcher had only a 10% chance of obtaining significant results.

Many believe the emphasis on type I error control used in popular procedures such as the analysis of variance follow up tests and the emphasis on teaching the easier concept of type I errors may have contributed to the lack of power we now see in statistical studies. One only needs to turn to the popular Dunn-Bonferroni, Scheffé, Tukey, and Newman-Keuls follow up procedures in the analysis of variance to see examples of attempts to stringently control type I errors. However, when type I errors are stringently controlled, the price that is paid is a lack of control of the inversely related type II error, lowered test power, and less chance of obtaining a significant result.

How much power do typical published studies have? Cohen (1962) was one of the first to point out the problem of low power when he reviewed 78 articles appearing in the 1960 *Journal of Abnormal and Social Psychology*. He found the mean power value of studies, assuming a medium effect size, was only .48, where effect size is the degree to which a phenomenon exists in a study. This finding indicated the researchers had slightly less than a 50 - 50 chance of rejecting the null hypothesis. For studies with small effects the odds were lower, and only when authors had large effects did they have a good chance, approximately 75%, of rejecting the null hypothesis.

With this information in hand, one would suspect researchers would be more cognizant of the power of their studies. However, when Sedlmeier and Gigerenzer (1989) replicated Cohen's study by reviewing 1984 articles, they found that the mean power of studies had actually declined from .48 to .37. It should be noted that Cohen's original methodology, used in these power studies, uses sample size and Cohen's definitions of large, medium, and small effects size to determine power rather than actual effect size (Thompson, 1998). As a result, the outcomes of these studies have been questioned. Nevertheless, they do point out the fact that decades of warnings about low power studies had done nothing to increase the power of studies.

One can only speculate on the damage to cumulative knowledge that has been cast upon the social sciences when study authors have only approximately a 50% chance of rejecting the null hypothesis and getting significant results. If the author does not obtain significant results in his/her study, the likelihood of being published is severely diminished due to the publication bias that exists for statistically significant results (Begg, 1994). As a result there may be literally thousands of studies with meaningful effect sizes that have been rejected for publication or never submitted for publication. These studies are lost because they do not pass muster with NHST. This is particularly problematic in educational research where effect sizes may be subtle but at the same time may indicate meritorious improvements in instruction and other classroom methods (Cohen, 1988).

Sample Size Dependence

The power of a statistical test, or how likely the test is to detect significant results, depends not only on the alpha and beta levels but also on the reliability of the data. Reliability is related to the dispersion or variability in the data, and as a result it can be controlled by reducing measurement and sampling error. However, the most common way of increasing reliability and increasing the power of a test is to increase the sample size.

With increased sample size, we incur yet another problem, that is the sample size dependency of tests used in NHST. Bakan (1966) reported on the results of a battery of tests he had collected on 60,000 subjects in all parts of the United States. When he conducted significance tests on these data, he found that every test yielded significant results. He noted that even arbitrary and nonsensical divisions, such as east of the Mississippi versus west of the Mississippi and Maine versus the rest of the country, gave significant results. "In some instances the differences in the sample means were quite small, but nonetheless, the *p* values were all very low" (p. 425). Nunnally (1960) reported similar results using correlation coefficients on 700 subjects and Berkson (1938) found similar problems using a chi-square test. Berkson stated, ". . . we have something here that is apt to trouble the conscience of a reflective statistician . . . a large sample is always better than a small sample . . . (and) . . . if we know in advance the *p* will result from . . . a test of a large sample . . . (then) . . . there would seem to be no use in doing it on a smaller one . . . since the result . . . is known, it is no test at all" (p. 526). Therefore, a small difference in estimates of population parameters from large samples, no matter how insignificant, yields significant results.

Ironically, if we have low test power, we cannot detect statistical significance, but if we have high test power, via a large sample size, all differences, no matter how small, are significant. Schmidt (1996a) has pointed out a troubling problem associated with solving power problems with large sample sizes. He suggested that scientific inquiry can be retarded because many worthwhile research projects cannot be conducted, since the sample sizes required to achieve adequate power may be difficult, if not impossible, to attain. It is not unusual for the educational researcher to have to settle for smaller samples than desired. Therefore, it is not likely that educational studies can escape the bane of low power as long as NHST is the statistical tool used. But before we worry too much about power problems in NHST, perhaps we should consider the thoughts of Oakes (1986) and later Schmidt (1996a). Schmidt noted that the power of studies "is a legitimate concept only within the context of statistical significance testing . . . (and) . . . if significance testing is no longer used, then the concept of statistical power has no place and is not meaningful" (p. 124).

Misunderstanding of p Values

With the advent of easy to use computer programs for statistical analysis, the researcher no longer has to depend on tables and the manual procedures for NHST, instead computerized statistical packages provide the researcher with a *p* value that is used to determine whether we reject, or fail to reject, the null hypothesis. As such, *p* values lower than the alpha value are viewed as a rejection of the null hypothesis, and *p* values equal to or greater than the alpha value are viewed as a failure to reject. The *p* value tells us nothing about the magnitude of significance nor does it tell us anything about the probability of replication of a study. The *p* value's use is limited to either rejecting or failing to reject the null hypothesis. It says nothing about the research or alternative hypothesis (Carver, 1978). The *p* value is primarily a function of effect size and sampling error (Carver, 1993). Therefore, differences of even trivial size can be judged to be statistically significant when sampling error is small (due to a large sample size and/or a large effect size) or when sampling error is large (due to a small sample size and/or a small effect size). However, NHST does not tell us what part of the significant differences is due to effect size and what part is due to sampling error.

The easy access to *p* values via statistical software has led in some instances to misunderstanding and misuse of this information. Since many researchers focus their research on *p* values, confusion about the meaning of a *p* value is often revealed in the literature. Carver (1978) and Thompson (1993), among others, have indicated that users of NHST often misinterpret the meaning of a *p* value as being a magnitude measure. This is evidenced by such common phrases, as "almost achieving significance" and "highly significant" (Carver, 1978, p. 386). They rightfully point out that many textbooks make the same mistake and that some textbooks have gone one step further by implying that a statistically significant *p* value indicates the probability that the results can be replicated. This is evidenced in statements such as "reliable difference" or the "results were reliable" (Carver, 1978, p. 385). No part of the logic of NHST implies this.

Thompson (1995, November) has noted that many researchers use the *p* value as a vehicle to "avoid judgment" (p. 10). He implies that when a significant result is obtained, the analyst is generally provided with the confidence to conclude his/her analysis. The devotion to *p* values to determine if a result is statistically significant suspends further analysis. Analysis should continue to determine if the statistically significant result is due to sampling error or due to effect size. For this information, the researcher will need to determine the effect size, using one of many available effect magnitude measures. He/she will then construct confidence intervals to assess the effect of sample size and error. As a last step, he/she will look to other methods to provide an indication of the replicability of the results. With this information in hand, the researcher can then not only better assess his/her results but can also provide more guidance to other researchers.

As this brief summary has shown, the simplicity and appeal of the dichotomous decision rule, posited by *p* values, is alluring. But, it can lead to misinterpretation of statistical significance, and more importantly it can distract us from a higher goal of scientific inquiry. That is, to determine if the results of a test have any practical value or not.

Defenders of NHST

With the plethora of shortcomings of NHST that have been documented for over 60 years, one would suspect there are few defenders of a procedure that suffers from so many weaknesses. In fact, Oakes (1986) has expressed, "It is extraordinarily difficult to find a statistician who argues explicitly in favor of retention of significance tests" (p. 71). Schmidt (1996a) reported that a few psychologists have argued in favor of retention of NHST, but "all such arguments have been found to be logically flawed and hence false" (p.116). As in all areas of endeavor, change is often difficult to accept, especially movement away from a phenomenon that has become an integral part of the work of so many people for so many years.

Winch and Campbell (1969), Frick (1996), and Cortina and Dunlap (1997) are among those who have spoken for the retention of significance testing. However, all of these defenders acknowledge the problematic nature and limited use of NHST. Winch and Campbell (1969), while defending NHST, stated, " . . . we advocate its use in a perspective that demotes it to a relatively minor role in the valid interpretation of . . . comparisons" (p. 140). The timidity of the typical defense was echoed by Levin (1993), when he stated, " . . . until something better comes along significance testing just might be science's best alternative" (p. 378).

With few strident defenders and almost universal detractors, the salient question is where do we go from here? Since our hallmark statistical test is flawed, do we have a replacement? We not only believe there is a replacement available now, but the replacement methods have the potential, if properly used, to move us out of the current morass described by Meehl (1978) more than 20 years ago. He described a situation in social sciences where theories are like fads. They come to the forefront with a flurry of enthusiasm, then they slowly fade away as both positive and negative results are gleaned from empirical data, and the results get more and more confusing and frustrating. This typical mixture of negative and positive findings is most likely the result of low power studies that sometimes reach statistical significance and sometimes do not.

Instead of all research effort contributing to the body of research knowledge, only the studies that are lucky enough to reach statistical significance via large sample size, or via chance, ever reach the research community. We would like to see a situation where all studies that were adequately designed, controlled, and measured would be reported, regardless of statistical significance. Below, we provide brief primers, along with appropriate references, to the tools that we believe will eventually replace the much flawed NHST.

Effect Magnitude Measures

In search of an alternative to NHST, methodologists have developed both measures of strength of association between the independent and dependent variables and measures of effect size. Combined, these two categories of measures are called "effect magnitude measures" (Maxwell & Delaney, 1990). Table 1 provides information on the known effect magnitude measures.

Table 1
Effect Magnitude Measures

Measures of Strength of Association	Measures of Effect Size
r, r_{pb}, R, R^2, η, η^2, η_{mult}	Cohen (1988) d, f, g, h,q, w
Cohen (1988) f^2	Glass (1976) g
Contingency coefficient	Hedges (1981) g
Cramer (1946) v	Tang (1938) ϕ
Fisher (1921) z	
Hays (1963) ω^2 and ρ_I	
Kelly (1935) ε^2	
Kendall (1963) W	
Tatsuoka (1973) $\hat{\omega}^2_{mult\ c}$	

Note. Eta squared (η^2) in ANOVA , called the correlation ratio, is the sum of squares (SS) for an effect divided by the SS_{total}. R^2 is the proportional reduction in error, or PRE, measure in regression. R^2 is the $SS_{regression}$ divided by SS_{total} Both η^2 and R^2 are analogous to the coefficient of determination (r^2). Adapted from Kirk, "Practical significance: A concept whose time has come." *Educational and Psychological Measurement*, *56*(5), p.749. Copyright 1996 by Sage Publication, Inc. Adapted with permission.

Measures of association are used for examining proportion of variance (Maxwell & Delaney, 1990, p. 98), or how much of the variability in the dependent variable(s) is associated with the variation in the independent variable(s). Common measures of association are the family of correlation coefficients (r), eta squared (η^2) in ANOVA, and R^2 (proportional reduction in error) in regression analysis.

Measures of effect size involve analyzing differences between means. Any mean difference index, estimated effect parameter indices, or standardized difference between means qualify as measures of effect size. It should be noted that effect size indices can be used with data from both correlational and experimental designs (Snyder & Lawson, 1993). Both measures of association and effect size can provide us with measures of practical significance when properly used.

Measures of Association

Kirk (1996) has reviewed the history of the development of these measures. Oddly, it was noted that Ronald Fisher, the father of NHST, was one of the first to suggest that researchers augment their tests of significance with measures of association (p. 748). Kirk found that effect magnitude measures other than the traditional measures of variance-accounted-for, such as r^2, are rarely found in the literature (p. 753). He believes this is due not to an awareness of the limitations of NHST but rather to the widespread use of regression and correlation procedures that are based on the correlation coefficient. However, the low instance of use of these measures could be due to their lack of availability in popular statistical software.

Snyder and Lawson (1993) have warned us of the perils of indiscriminate use of measures of association. They indicate that experimental studies and more homogeneous samples result in smaller measures of association and that studies that involve subject-to-variable ratios of 5:1 or less will usually contain noteworthy positive bias (p. 339). Issues such as the study design (fixed or random effects designs) and whether we are using univariate or multivariate measures also impact the choice of measure of association. In general, formulas designed to estimate measures of association in other samples are less biased than formulas designed for estimating measures of association in the population. Also, a study that has a large effect size and a large sample size will typically need no correction for bias, however smaller effect sizes and smaller sample sizes should use measures corrected for positive bias. For a detailed explanation of appropriate measures of association as well as computational formulas, the reader is referred to either Snyder and Lawson (1993) or Maxwell and Delaney (1990). Various measures of association are shown in Table 1.

Measures of Effect Size

Perhaps no one has done more than Jacob Cohen to make researchers aware of the use of effect size measures, as well as the problem of low test power in NHST. Cohen (1988) also provides us with definitions of effect size as well as conventions that can be used in the absence of specific information regarding a phenomenon. The various effect size measures are outlined in Table 1. Effect size is defined "without any necessary implication of causality . . . (as) . . . the degree to which the phenomenon is present in the population . . . or . . . the degree to which the null hypothesis is false" (p. 9). Cohen further states, "the null hypothesis always means the effect size is zero" (p. 10). A generalized form of effect size *d* is used for independent samples in a one-tailed, directional case:

$$d = \mu_1 - \mu_2 / \sigma$$

where *d* is the effect size index for the *t* test for means, μ_1 and μ_2 are population means, and σ is the population standard deviation. As such, the value of the difference in the population means is divided by the population standard deviation to yield a standardized, scale invariant, or metric-free, estimate of the size of the effect.

Substituting sample statistics in the formula as estimates of the population parameters can also be applied. The standard deviation can either be the standard deviation of a control group, assuming equality of variance, or alternatively the pooled (within) population standard deviation can be used (Wolf, 1986). Cohen has developed methods of converting most of the popular significance tests to effect size measures. For example, there are effect size measures for differences between correlation coefficients (q), differences between proportions (h), the chi-square test for goodness of fit and contingency tables (w), ANOVA and ANCOVA (f), multiple regression and other multivariate methods (f^2). The reader is referred to Cohen (1988) for a full treatment of this subject.

Interpreting Effect Size

Various interpretation methods have been developed for effect size measures. Cohen (1988) developed three measures of overlap or U measures. With the assumptions of normality and equality of variance satisfied, and with two populations, A and B, U_1 is defined as the percentage of combined area not shared by the two populations distributions. U_2 is the percentage in the B population that exceeds the same percentage in the A population. U_3 is the percentage of the A population which the upper half of the cases of the B population exceeds. Cohen provides tables to determine the U measures for effect sizes 0 - 4 (p. 22). The U_3 measure of overlap can be interpreted using the tabled values of the standard normal distribution. For example, if effect size, *d*, is .5 (a medium effect), the area under the normal curve would be .6915 (.5 +.1915). This means that the treatment effect would be expected to move a typical person from the 50th percentile to the 69th percentile of the control group. Generally, the result of this outcome is graphically displayed for easier interpretation. The reader is referred to Glass (1976) for one of the earliest uses of this interpretive device.

Rosenthal and Rubin (1982) have described a method for evaluating the practical significance of the effect size measures that has shown promise. This procedure transforms *r*, or other effect measures, to chi-square (χ^2) to form a binomial effect size display (BESD) for 2 x 2 tables. The relatively easy calculations provide us with the estimated difference in success probabilities between the treatment and control groups. This method holds promise, but criticism has surfaced that attacks the method as distorting the data (McGraw, 1991), especially in cases where differences are highly divergent from 50-50 (Strahan, 1991), and as misinterpreting the data (Crow, 1991). Rosenthal (1991) has responded by noting that this method is context specific and was not intended to assess all situations. As a result, caution should be exercised when using BESD tables, especially in cases where differences in treatment and control groups are large.

Interpretation of the effect size is best accomplished by comparing the study effect size to the effect size of similar studies in the field of study. Methods for determining a general effect size in a particular field of study have been limited to studies of the *median* effect size of studies in a particular journal (Haase, Waechter, & Solomon, 1982). This type of study converts traditional test statistics into a distribution of effect sizes and provides a convenient method of comparing results of a single test to that of results in the field as a whole. We believe more studies of this type, along with periodic updates, would provide the primary researcher with the most valid assessment of a particular effect size. In lieu of this type of information, Cohen (1988) has provided general conventions for the use of effect size. A small effect is defined as .2, a medium effect as .5, and a large effect as .8. Cohen warns that these conventions are analogous to the conventions for significance levels (α = .05) and should be used with great caution, and only in the case where previous research is unavailable (p. 12). However, Kirk (1996) has noted that the average effect size of observed effects in many fields approximates .5 and the meaning of effect size remains the same without regard to the effect size measure. In general, the ultimate judgment regarding the significance of the effect size measure "rests with the researcher's personal value system, the research questions posed, societal concerns and the design of a particular study" (Snyder & Lawson, 1993, p. 347). Both Snyder and Lawson (1993) and Thompson (1993a, pp. 365-368) provide very readable information on the calculation, as well as the use and limitations of univariate and multivariate effect magnitude measures.

Confidence Intervals

The traditional NHST provides us only with information about whether chance is or is not an explanation for the observed differences. Typically, the use of confidence intervals is treated as an alternative to NHST since both methods provide the same outcome. Point estimates of differences, surrounded by confidence intervals, provide all the information that NHST does, but additionally they provide the degree of precision observed, while requiring no more data than NHST. Surprisingly, based on a review of recent literature, the superiority of this method is not recognized or has been ignored by the research community (Kirk, 1996, p. 755). Why should we routinely report confidence intervals? Not only do they serve to remind the researcher of the error in his/her results and the need to improve measurement and sampling techniques, they also provide a basis for assessing the impact of sample size. Note that confidence intervals are an analogue for test power. A larger sample size, higher power test will have a smaller confidence interval, while a smaller sample size, lower power test will have a larger confidence interval.

Work on asymmetric confidence intervals and expanding the use of confidence intervals to apply to multivariate techniques and causal models has been underway for some time. Many of the methods have been available but were so complex that they were seldom used. However, the use of high speed computers makes calculations of these confidence intervals more realistic. A detailed look at more recent and appropriate applications of confidence intervals have been described by Reichardt and Gollob (1997) and Serlin (1993).

In summary, there is a multitude of effect magnitude measures available to provide the practical significance of effects revealed in a study. When used in combination with confidence intervals that describe sampling error, magnitude measures present the researcher with more information than is provided by NHST. However, the use of these measures has not yet received widespread acceptance by the research community. We believe the lack of acceptance is due not to active resistance but to a lack of familiarity with effect magnitude measures and confidence intervals when compared with NHST. Some may argue that the interpretation of these measures is more subjective than the dichotomous interpretation of significance tests. However, those arguments fail to consider the subjectivity of the significance level in NHST and the general subjective nature of all empirical science (Thompson, 1993).

Simulated Replications

Fisher (1971), among others, has acknowledged the need for replication of studies in order to verify results and, in the current vernacular, to advance cumulative knowledge. However, there are many factors working against replication studies. Among them are a general disdain for non-original research by journal editors and dissertation committees, lack of information on another's study to replicate it, and the bias that is implied when the researcher replicates his/her own study. Additionally, replication of one's own study immediately following its completion is likely to invoke a strong fatigue factor. Nevertheless, some indication of the likelihood of replicability of results is in the interest of good science.

Fortunately, there are alternatives to full-scale replication. Schmidt (1996a) has noted that the power of a test provides us with an estimate of the probability of replication (p.125), and Thompson (1993a) describes three methods that can be used to indicate the likelihood of replication. Two of the methods, crossvalidation and the jackknife techniques, use split samples to empirically compare results across the sample splits. The third method, bootstrapping, involves sampling equal size samples with replacement from the original data set. After several thousand iterations, one is provided with an analogue to the sampling distribution of means. The resulting data have a variety of uses including estimating the standard error of the means, developing confidence intervals around the estimate of the population mean, and providing a vehicle for viewing the skewness and kurtosis in a simulated population distribution. Thompson pointed out two practical uses of the bootstrap method: 1) to descriptively evaluate the stability of the results of the study, and 2) to make inferences using confidence intervals (p. 372). Statistical software designed by researchers for the specific purpose of conducting bootstrap studies are available (p. 369). The one thing the researcher should always consider when conducting a bootstrap study is the inherent limitations of the original data that are carried over to the bootstrap method. As a result, caution and thoughtfulness in the interpretation of data are called for in this, as in all statistical analyses. In summary, the reporting of studies should include some indication of the replicability of the data. No matter what method the author chooses, it will provide more information than is available from NHST.

Meta-analysis

Meta-analysis is defined as, ". . . the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). In the past, subjective literature reviews or simplistic vote counting of significant and non-significant results were used. Light and Pillemer (1984) described these methods as subjective, scientifically unsound, and an inefficient way to extract useful information. Cooper and Hedges (1994) describing the early meta-analyses stated, "research synthesis in the 1960s was at best an art, at worst a form of yellow journalism" (p. 7). However, the field of meta-analysis has seen a burst of activity since Glass (1976) first coined the term and used Cohen's effect size and overlap measures to analyze psychotherapy outcome research. Glass paved the way for a plethora of meta-analytic studies in the 1980s and 1990s that used effect size as the dependent variable. Cooper and Hedges (1994) observed that "much of the power and flexibility of quantitative research synthesis is owed to the existence of effect size estimators such as r and d" (p. 24). The power of these statistics comes from their ability to measure the effects in terms of their own standard deviations.

With the advances in the development of effect size measures and meta-analytic techniques, the field of meta-analysis now has a body of statistics specifically for combining the results of studies (Hedges & Olkin, 1985). Additionally, many of the early methods of meta-analysis have been "standardized" and many of the early criticisms of meta-analysis (Wittrock, 1986) have been addressed (Cooper & Hedges, 1994). Today, we see the role of meta-analysis taking on more and more importance in scientific inquiry. This is evidenced by a growing number of meta-analytic studies published in journals that formerly refused to publish literature reviews, as well as shifting patterns of citations in the literature (Schmidt, 1996a). In a recent development, meta-analytic methods have now been broadened to the empirical study of variability of test score reliability coefficients across samples. This reliability generalization method along with extant validity generalization methods makes meta-analysis an even more powerful method of data synthesis (Vacha-Haase, 1998). The interested reader should consult Cooper and Hedges' (1994) text on methods, statistics and limitations of current meta-analytic practices. The development of meta-analysis as an "independent specialty within the statistical sciences" (p. 6) allows the secondary researcher to use sound statistical methods to combine the results of years of research to interpret a phenomena.

Research Registries

Despite the fact that many of the methods of meta-analysis come from the social sciences, the more dramatic use of these methods has been in the field of health care. This development was most likely due to the availability of registries of studies in the health care field. By tracking all known research studies in specialty areas, the field had a wealth of data to draw upon. Meta-analysis has been so successful in medical research that federal legislation has authorized creation of an agency for health care policy research that is required to develop guidelines based on a systematic synthesis of research evidence (Cooper & Hedges, 1994, p. 7).

One of the problems facing the registries in health care is lack of knowledge in the field about their availability. There are so many registries for so many clinical trials that registries of registries have had to be formed. In the social sciences we can learn a lesson from the ad hoc nature of establishing registries that has developed in medical science. Dickersin (1994) notes that the institutional review system for research registration already exists for all research involving human subjects. She has identified a national system that exists in Spain that mandates cooperation between local institutional review boards and a centralized national board (p. 71). With the availability of high speed electronic transfer of data, what would have seemed like a pipe dream some years ago now has the possibility of becoming a reality. A national system for the social sciences, working through local review boards, could be stimulated through concerted action by a coalition of professional organizations and the federal government. However, if government intervention is unthinkable, perhaps professional organizations could muster the manpower and resources to develop research registries in education and/or psychology.

Where We Go from Here

Based on our review of the arguments and logic of NHST and the vast literature on augmentation and replacement methods, we have come to the conclusion (albeit not a unique or new conclusion) that individual studies can best be analyzed by using point estimates of effect size as a measure of the magnitude of effect and confidence limits as a measure of the sampling error. Reporting these findings will provide more detailed information and certainly more raw information than is contained in significance tests (Schafer, 1993). Additionally, individual studies should indicate the likelihood of replication through the use of simulation methods. The researchers who believe the *p* value provides this information are thinking appropriately, but incorrectly, in that replication is the only way to reach consensus on the evidence provided by individual studies. However, statistical tools that simulate replications are the best methods of providing evidence of replicability, short of full-scale replication. We also believe the academic community should rethink the importance and the role of full-scale replication studies in scientific investigation and promote them to a status equal to that of original research. These recommendations should bring some order to the chaotic situation that currently exists in the analysis of individual studies. Using the described methods and with the availability of research registries, the meta-analytic researcher will have access to more studies (including those formerly unsubmitted or rejected as non-significant), and the studies will be reported in a manner that is more conducive to meta-analytic studies.

We believe a major advancement of knowledge will come from a synthesis of many individual studies regarding a particular phenomenon using meta-analytic methods. With the primary researcher providing raw materials, the meta-analytic secondary researcher can analyze trends in various areas of research endeavor and provide the raw materials for more rational educational policy.

Changing Times

There are signs that the mountain of criticism that has befallen NHST has finally reached fruition. There is evidence in the research environment that change is taking place and the abandonment of NHST for the use of point estimates of effect size with confidence intervals is underway. In 1996, the American Psychological Association's Board of Scientific Affairs formed a task force to study and make recommendations about the conduct of data analysis (APA Monitor, 1997). The initial report of the committee fell short of recommending a ban on NHST, however it did report that "... (data analysis)... include both direction and size of effect and their confidence intervals be provided routinely . . ." (APA Science Agenda, 1997, p. 9). Two years earlier, and almost unnoticed, the fourth edition of the APA Publication Manual (1994) stated, "You are encouraged to provide effect-size information. . . whenever test statistics and samples sizes are reported" (p. 18). Kirk (1996) reported the APA is also seeking involvement from the AERA, APS, Division 5, the Society for Mathematical Psychology and the American Statistical Association in its study of the NHST issue (p. 756). Schmidt (1996a) reported that studies today are more likely to report effect sizes, and "it is rare today in industrial/organizational psychology for a finding to be touted as important solely on the basis of its *p* value" (p. 127). Additionally, government entities are now seeing the importance of meta-analytic studies and the effect size measures they use and are calling for more studies to guide policy decisions (Sroufe, 1997). Popular statistical software is also being reprogrammed to provide measures of power and effect size (J. McLean, personal communication, November 12, 1997).

Despite the fact that Michigan State has reformed its graduate statistics course sequence in psychology to include teaching of effect size measures and a de-emphasis of NHST (Schmidt, 1996a), it is acknowledged that "there have been no similar improvements in the teaching of quantitative methods in graduate and undergraduate programs" (p. 127). This mirrors a report (Aiken, West, Secrest, & Reno, 1990) that reviewed Ph.D. programs in psychology and concluded that "the statistics . . . curriculum has advanced little in 20 years" (p. 721). Thompson (1995) has also noted that his review of AERA publications and of papers presented at (the) annual meetings suggest that the calls for new methods haven't affected contemporary practice. Based on our own knowledge of teaching methods and statistics textbooks, we do not believe the academic community or textbook publishers have changed appreciably since the 1990 report issued by Aiken, et al. (1990).

Strategies for Change

We respect democratic principles so we cannot in good faith call for a ban on significance testing since this would represent censorship and infringement on individual freedoms. However, we believe that most statisticians would welcome orderly change that would lead to abandonment of NHST. In no way would it prohibit the diehard researcher from using NHST, but all emphasis would be on improved methods of legitimate research. These methods would be directed at ways and means of facilitating meta-analytic studies. This would include editorial policies that require: a) validity and reliability measures on all instruments used; b) use of appropriate effect magnitude measures with confidence intervals to describe studies; c) use of information such as effect size studies of the phenomena of interest, BESD methods, odds ratio's, Cohen's effect size interpretations and other measures to interpret the results; and d) an indication of the replicability of the results obtained using bootstrap or other legitimate methods. Educational research registries would be put in place to attempt to replicate the registries that have demonstrated success in the health care field. Statistical software would be modified to emphasize the procedures and caveats for the newer statistical methods (including meta-analysis), and textbooks would be revised to reflect the changes in emphasis.

We see the various stakeholders, or interest groups, in the discussion we have presented as: a) professional associations, b) journal editors, c) researchers, d) educators, e) statistics textbook writers, and f) statistical software developers. The first steps in replacing NHST have taken place with professional organizations addressing the issue of NHST. We believe this step will eventually influence editorial policies used by journal editors. This, we believe, will be the critical path for change since it will, in turn, influence the researchers' data analyses and writings, as well as their educational practices.

For the above scenario to occur with minimal disruption, a joint project of the leading professional organizations needs to take the first step with a well developed master plan for change. Prominent practitioners, not dissimilar from the extant APA task force on significance testing, would outline a general framework for change following suggestions outlined in this and other works that have taken a critical look at the issues surrounding current research practice.

Following the development of the general plan, several other task forces of prominent practitioners would be formed to flesh out the details for the master plan. We envision these task forces addressing the issues of editorial policies for scholarly journals, revisions required to be made by textbook and statistical software publishers, and development of research registries. Once the individual task forces had reported, their work would be put out for review and comment by the interested professionals.

The original master plan task force would coordinate the final development of the master plan, based on the input of the various task forces and the public comment. The professional organization would then announce the date for the change-over that would give all stakeholders time to prepare. An analogy would be the rollout of a new computer operating system, where software developers, vendors and users are aware of and prepared for the change that is going to take place long before it actually occurs. Users are kept aware of the progress of change through periodic, well publicized and distributed information. This process would allow an orderly and expedited process. We would envision the above described process entailing approximately 24 to 36 months of concerted effort.

Summary

With the evidence that has been provided, it is reasonable to state that NHST, with its many shortcomings, has failed in its quest to move the social sciences toward verisimilitude and may have actually stymied the advancement of knowledge. NHST promised an improved method of determining the significance of a study, and no doubt was enlightening in the 1930s when researchers were saddled with fewer methods of inquiry. Some sixty years later, we can now state that methods with the track record of NHST have no place in scientific inquiry. In the past, we may have had to tolerate the shortcomings of NHST because there were no viable alternatives. Today viable and continually evolving alternatives are available. The use of effect magnitude measures, replication measures, and the statistics that drive meta-analytic studies are no longer embryonic, and we believe they merit a central role in scientific inquiry.

The loss of NHST techniques will not mean that older studies are meaningless. In fact, many studies that have failed to pass the NHST test and were not published or presented can be resurrected and updated with effect size measures. As a result, the loss of NHST will not retard the growth of scientific knowledge but will, ironically, advance scientific knowledge. We strongly believe a major step in advancing cumulative knowledge will be the establishment of research registries to compile all studies of a particular phenomenon for meta-analysis. Controversy will always surround statistical studies, and this paper in no way proposes that current effect magnitude measures and meta-analytic techniques are without limitations. We will see misuses of the measures that we propose, just as we have seen misuses of NHST, but we should remain vigilant and not allow these misuses to be institutionalized as they apparently have been with NHST. With change, the new century promises more advanced and enlightened methods will be available to help forge more rational public policies and advance the cumulative knowledge of educational research, in particular, and the social sciences, in general.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. L. (1990). Graduate training in statistics, methodology, and measurement in psychology, a survey of Ph.D. programs in North America. *American Psychologist, 45*(6), 721-734.

APA Monitor. (1997, March). *APA task force urges a harder look at data, 28*(3), 26. Washington, D.C.: Author.

APA Science Agenda (1997, March-April). *Task force on statistical inference identifies charge and produces report, 10*(2), 9-10. Washington, D.C.: Author.

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, D.C.: Author.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423-437.

Begg, C. B., (1994). Publication bias. In H. Cooper & L. V. Hedges, (Eds.) *The Handbook of Research Synthesis.* (pp. 399-409). New York: Russell Sage Foundation.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526-542.

Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review, 48,* 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

Cohen, J. (1962). The statistical power of abnormal social psychology research. *Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences,* (2nd ed.). Hillsdale, N.J.; Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*(12), 997-1003.

Cooper, H. M. & Hedges, L. V. (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cortina, J. M. & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*(2), 161-172.

Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?" *American Psychologist, 46,* 1083.

Dickersin, K. (1994). Research registries. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 71). New York: Russell Sage Foundation.

Fisher, R. A. (1971). *The design of experiments.* (8th ed.) New York: Hafner Publishing.

Frick, R. W. (1996) The appropriate use of null hypothesis testing. *Psychological Methods, 1*(4), 379-390.

Glass, G. V. (1976). Primary, secondary and meta-analysis. *Educational Researcher, 5,* 3-8.

Haase, R., Waechter, D., & Solomon, G. (1982). How significant is a significant difference? Average effect size of research in counseling. *Journal of Counseling Psychology, 29,* 58-65.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego; Academic Press.

Hinkle, D. E., Wiersma, W., & Jurs, S.G. (1994). *Applied statistics for the behavioral sciences* (3rd ed.). Boston; Houghton Mifflin Company.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*(4), 317-333.

Jones, L. V. (1955). Statistical theory and research design. *Annual Review of Psychology, 6,* 405-430.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*(4), 378-381.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36*(2), 102-105.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth Publishing.

McGraw, K. O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist, 46*, 1084-1086.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.

Morrison, D. E., & Henkel, R. E. (1970). *The significance testing controversy - A reader*. Chicago: Aldine Publishing.

Mulaik, S. A, Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Lawrence Erlbaum Associates.

Nunnally, J. (1960). The place of statistics in psychology. *Education and Psychological Measurement, 20*, 641-650.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, John Wiley & Sons.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259-284). Mahwah, N.J.: Lawrence Erlbaum Associates.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD and alternative indices. *American Psychologist, 46*, 1086-1087.

Rosenthal, R., & Rubin, D., (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp.176-197). Mahwah, NJ: Lawrence Erlbaum Associates.

Rozeboom, W. W. (1960). The fallacy of null hypothesis significance testing. *Psychological Bulletin, 57*, 416-428.

Schafer, J. P. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education, 61*(4), 383-387.

Schmidt, F. L. (1996a). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115-129.

Schmidt, F. L. (1996b). What do data really mean? Research findings, meta analysis and cumulative knowledge in psychology. *American Psychologist, 47*(10), 1173-1181.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105* (2), 309-316.

Serlin, R. C. (1993). Confidence intervals and the scientific method: Case for Holm on the range. *Journal of Experimental Education, 61*(4), 350-360.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*(4), 334-349.

Sroufe, G. E. (1997). Improving the "awful reputation" of educational research. *Educational Researcher, 26*(7), 26-28.

Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist, 46*, 1083-1084.

Thompson, B. (1993a). Foreword. *Journal of Experimental Education, 61*(4), 285-286.

Thompson, B. (1993b). The use of significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, 6(4), 361-377.

Thompson, B. (1995). *Inappropriate statistical practices in counseling research: Three pointers for readers of research literature*. Washington, D. C. Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. 391 990).

Thompson, B. (1995, November). *Editorial policies regarding statistical significance testing: Three suggested reforms*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Thompson, B. (1998). [Review of the book *What if there were no significance tests?*] *Educational and Psychological Measurement*, (in press).

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20.

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence, yes. The significance of tests of significance. *American Sociologist, 4*, 140-143.

Wittrock, M. C. (1986). *Handbook of Research on Teaching*, (3rd ed.). New York: MacMillan Publishing.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*, (Series no. 07-059). Newbury Park, CA: Sage Publications.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 15-22

The Role of Statistical Significance Testing In Educational Research

James E. McLean
University of Alabama at Birmingham

James M. Ernest
State University of New York at Buffalo

The research methodology literature in recent years has included a full frontal assault on statistical significance testing. The purpose of this paper is to promote the position that, while significance testing as the sole basis for result interpretation is a fundamentally flawed practice, significance tests can be useful as one of several elements in a comprehensive interpretation of data. Specifically, statistical significance is but one of three criteria that must be demonstrated to establish a position empirically. Statistical significance merely provides evidence that an event did not happen by chance. However, it provides no information about the meaningfulness (practical significance) of an event or if the result is replicable. Thus, we support other researchers who recommend that statistical significance testing must be accompanied by judgments of the event's practical significance and replicability.

The research methodology literature in recent years has included a full frontal assault on statistical significance testing. An entire edition of a recent issue of *Experimental Education* (Thompson, 1993b) explored this controversy. There are some who recommend the total abandonment of statistical significance testing as a research methodology option, while others choose to ignore the controversy and use significance testing following traditional practice. The purpose of this paper is to promote the position that while significance testing by itself may be flawed, it has not outlived its usefulness. However, it must be considered in the total context of the situation. Specifically, we support the position that statistical significance is but one of several criteria that must be demonstrated to establish a position empirically. Statistical significance merely provides evidence that an event did not happen by chance. However, it provides no information about the meaningfulness (practical significance) of an event or if the result is replicable.

This paper addresses the controversy by first providing a critical review of the literature. Following the review are our summary and recommendations. While none of the recommendations by themselves are entirely new, they provide a broad perspective on the controversy and provide practical guidance for researchers employing statistical significance testing in their work.

Review of the Literature

Scholars have used statistical testing for research purposes since the early 1700s (Huberty, 1993). In the past 300 years, applications of statistical testing have advanced considerably, most noticeably with the advent of the computer and recent technological advances. However, much of today's statistical testing is based on the same logic used in the first statistical tests and advanced in the early twentieth century through the work of Fisher, Neyman, and the Pearson family (see the appendix to Mulaik, Raju, & Harshman, 1997, for further information). Specifically, significance testing and hypothesis testing have remained at the cornerstone of research papers and the teaching of introductory statistics courses. (It should be noted that while the authors recognize the importance of Bayesian testing for statistical significance, it will not be discussed, as it falls outside the context of this paper.) Both methods of testing hold at their core basic premises concerning probability. In what may be termed Fisher's *p value approach*, after stating a null hypothesis and then obtaining sample results (i.e., "statistics"), the probability of the sample results (or sample results more extreme in their deviation from the null) is computed, assuming that the null is true in the population from which the sample was derived (see Cohen, 1994 or Thompson, 1996 for further explanation). The Neyman-Pearson or *fixed-alpha approach* specifies a level at which the test statistic should be rejected and is set a priori to conducting the test of data. A null hypothesis (H_o) and an alternative hypothesis (H_a) are stated, and

---

James E. McLean is a university research professor and the director of the Center for Educational Accountability in the School of Education at the University of Alabama at Birmingham. James M. Ernest is a lecturer in the Department of Learning and Instruction, Graduate School of Education, State University of New York at Buffalo. Correspondence relevant to this article should be addressed to James E. McLean, Center for Educational Accountability, University of Alabama at Birmingham, 901 13th Street South, Birmingham, AL 35294-1250 or by e-mail to jmclean@uab.edu.

if the value of the test statistic falls in the rejection region the null hypothesis is rejected in favor of the alternate hypothesis. Otherwise the null hypothesis is retained on the basis that there is insufficient evidence to reject it.

Distinguishing between the two methods of statistical testing is important in terms of how methods of statistical analysis have developed in the recent past. Fisher's legacy of statistical analysis approaches (including ANOVA methods) relies on subjective judgments concerning differences between and within groups, using probability levels to determine which results are statistically significant from each other. Karl Pearson's legacy involves the development of correlational analyses and providing indexes of association. It is because of different approaches to analyses and different philosophical beliefs that the issue of testing for statistical significance has risen. In Huberty's (1993) historical review of the importance of statistical significance testing literature, the research community has shifted from one perspective to another, often within the same article. Currently we are in an era where the value of statistical significance testing is being challenged by many researchers. Both positions (arguing for and against the use of statistical significance tests in research) are presented in this literature review, followed by a justification for our position on the use of statistical significance testing as part of a comprehensive approach.

As previously noted, the research methodology literature in recent years has included a full frontal assault on statistical significance testing. Of note, an entire edition of *Experimental Education* explored this controversy (Thompson, 1993b). An article was written for *Measurement and Evaluation in Counseling and Development* (Thompson, 1989). The lead section of the January, 1997 issue of *Psychological Science* was devoted to a series of articles on this controversy (cf., Hunter, 1997). An article suggesting editorial policy reforms was written for the American Educational Research Association (Thompson, 1996), reflected on (Robinson & Levin, 1997), and a rejoinder written (Thompson, 1997). Additionally, the American Psychological Association created a Task Force on Statistical Inference (Shea, 1996), which drafted an initial Report to the Board of Scientific Affairs in December 1996, and has written policy statements in the *Monitor*.

The assault is based on whether or not statistical significance testing has value in answering a research question posed by the investigators. As Harris (1991) noted, "There is a long and honorable tradition of blistering attacks on the role of statistical significance testing in the behavioral sciences, a tradition reminiscent of knights in shining armor bravely marching off, one by one, to slay a rather large and stubborn dragon . . . . Given the cogency, vehemence and repetition of such attacks, it is surprising to see that the dragon will not stay dead" (p. 375). In fact, null hypothesis testing still dominates the social sciences (Loftus & Masson, 1994) and still draws derogatory statements concerning the researcher's methodological competence. As Falk and Greenbaum (1995) and Weitzman (1984) noted, the researchers' use of the null may be attributed to the experimenters' ignorance, misunderstanding, laziness, or adherence to tradition. Carver (1993) agreed with the tenets of the previous statement and concluded that "the best research articles are those that include *no* tests of statistical significance" (p. 289, italics in original). One may even concur with Cronbach's (1975) statement concerning periodic efforts to "exorcize the null hypothesis" (p. 124) because of its harmful nature. It has also been suggested by Thompson, in his paper on the etiology of researcher resistance to changing practices (1998, January) that researchers are slow to adopt approaches in which they were not trained originally.

In response to the often voracious attacks on significance testing, the American Psychological Association, as one of the leading research forces in the social sciences, has reacted with a cautionary tone: "*An APA task force won't recommend a ban on significance testing, but is urging psychologists to take a closer look at their data*" (Azar, 1997, italics in original). In reviewing the many publications that offer advice on the use or misuse of statistical significance testing or plea for abstinence from statistical significance testing, we found the following main arguments for and against its use: (a) what statistical significance testing does and does not tell us, (b) emphasizing effect-size interpretations, (c) result replicability, (d) importance of the statistic as it relates to sample size, (e) the use of language in describing results, and (f) the recognition of the importance of other types of information such as Type II errors, power analysis, and confidence intervals.

What Statistical Significance Testing Does and Does Not Tell Us

Carver (1978) provided a critique against statistical significance testing and noted that, with all of the criticisms against tests of statistical significance, there appeared to be little change in research practices. Fifteen years later, the arguments delivered by Carver (1993) in the *Journal of Experimental Education* focused on the negative aspects of significance testing and offered a series of ways to minimize the importance of statistical significance testing. His article indicted the research community for reporting significant differences when the results may be trivial, and called for the use of effect size estimates and study replicability. Carver's argument focused on what statistical significance testing *does not do*, and proceeded to highlight ways to provide indices of practical significance and result replicability. Carver (1993) recognized that 15 years of trying to extinguish the use of statistical significance testing has resulted in little change in the use and frequency of statistical significance testing. Therefore the tone of the 1993 article differed from the 1978 article in shifting from a dogmatic anti-statistically significant approach to more of a bipartisan approach where the limits of significance testing were noted and ways to decrease their influence provided. Specifically, Carver (1993) offered four ways to minimize the importance of statistical significance testing: (a) insist on the word *statistically* being placed in front of significance testing, (b) insist that the results always be interpreted with respect to the data first, and statistical significance second, (c) insist on considering effect sizes (whether significant or not), and (d) require journal editors to publicize their views on the issue of statistical significance testing prior to their selection as editors.

Shaver (1993), in the same issue of *The Journal of Experimental Education*, provided a description of what significance testing is and a list of the assumptions involved in statistical significance testing. In the course of the paper, Shaver methodically stressed the importance of the assumptions of random selection of subjects and their random assignment to groups. Levin (1993) agreed with the importance of meeting basic statistical assumptions, but pointed out a fundamental distinction between statistical significance testing and statistics that provide estimates of practical significance. Levin observed that a statistically significant difference gives information about *whether* a difference exists. As Levin noted, if the null hypothesis is rejected, the *p* level provides an "a posteriori indication of the probability of obtaining the outcomes as extreme or more extreme than the one obtained, given the null hypothesis is true" (p. 378). The effect size gives an estimate of the noteworthiness of the results. Levin made the distinction that the effect size may be necessary to obtain the size of the effect; however, it is statistical significance that provides information which alludes to whether the results may have occurred by chance. In essence, Levin's argument was for the two types of significance being complementary and not competing concepts. Frick (in press) agreed with Levin: "When the goal is to make a claim about how scores were produced, statistical testing is still needed, to address the possibility of an observed pattern in the data being caused just by chance fluctuation" (in press). Frick's thesis concerning the utility of the statistical significance test was provided with a hypothetical situation in mind: the researcher is provided with two samples who together are the population under study. The researcher wants to know whether a particular method of learning to read is better than another method. As Frick (in press) noted,

> statistical testing is needed, despite complete knowledge of the population. The . . . experimenter wants to know if Method A is better than Method B, not whether the population of people learning with Method A is better than the population of people learning with Method B. The first issue is whether this difference could have been caused by chance, which is addressed with statistical testing. The example is imaginary, but a possible real-life analog would be a study of all the remaining speakers of a dying language, or a study of all of the split-brain patients in the world.

One of the most important emphases in criticisms of contemporary practices is that researchers must evaluate the practical importance of results, and not only statistical significance. Thus, Kirk (1996) agreed that statistical significance testing was a necessary part of a statistical analysis. However, he asserted that the time had come to include practical significance in the results. In arguing for the use of statistical significance as necessary, but insufficient for interpreting research, Suen (1992) used an 'overbearing guest' analogy to describe the current state of statistical significance testing. In Suen's analogy, statistical significance is the overbearing guest at a dinner party who

> inappropriately dominates the activities and conversation to the point that we forget who the host was. We cannot disinvite this guest. Instead, we need to put this guest in the proper place; namely as one of the many guests and by no means the host. (p. 78)

Suen's reference to a "proper place" is a call for researchers to observe statistical significance testing as a means to "filter out the sampling fluctuations hypothesis so that the observed information (difference, correlation) becomes slightly more clear and defined" (p. 79). The other "guests" that researchers should elevate to a higher level include ensuring the quality of the research design, measurement reliability, treatment fidelity, and using sound clinical judgment of effect size.

For Frick (in press), Kirk (1996), Levin (1993), and Suen (1992), the rationale for statistical significance testing is independent of and complementary to tests of practical significance. Each of the tests provides distinct pieces of information, and all three authors recommend the use of statistical significance testing; however, it must be considered in combination with other criteria. Specifically, statistical significance is but one of three criteria that must be demonstrated to establish a position empirically (the other two being practical significance and replicability).

Emphasizing Effect-Size Interpretations

The recent American Psychological Association (1994) style manual noted that

> Neither of the two types of probability values [statistical significance tests] reflects the importance or magnitude of an effect because both depend on sample size . . . You are [therefore] *encouraged* to provide effect-size information. (p. 18, italics added)

Most regrettably, however, empirical studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "*encouraging*" effect size reporting (American Psychological Association, 1994) has *not* appreciably affected actual reporting practices (e.g., Kirk, 1996; Snyder & Thompson, in press; Thompson & Snyder, 1997, in press; Vacha-Haase & Nilsson, in press). Due to this lack of change, authors have voiced stronger opinions concerning the "emphasized" recommendation. For example, Thompson (1996) stated "AERA should venture beyond APA, and *require* such [effect size] reports in all quantitative studies" (p. 29, italics in original).

In reviewing the literature, the authors were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report. Huberty (1993) noted that "of course, empirical researchers should not rely exclusively on statistical significance to assess results of statistical tests. Some type of measurement of magnitude or importance of the effects should also be made" (p. 329). Carver's third recommendation (mentioned previously) was the inclusion of terms that denote an effect size measure; Shaver (1993) believed that "studies should be published without tests of statistical significance, but not without effect sizes" (p. 311); and Snyder and Lawson (1993) contributed a paper to *The Journal of Experimental Education* special edition on statistical significance testing titled "Evaluating Results Using Corrected and Uncorrected Effect Size Estimates." Thompson (1987, 1989, 1993a, 1996, 1997) argued for effect sizes as one of his three recommendations (the language use of statistical significance and the inclusion of result replicability results were the other two); Levin (1993) reminded us that "statistical significance (alpha and p values) and practical significance (effect sizes) are not *competing* concepts– they are *complementary* ones" (p.379, italics in original), and the articles by Cortina and Dunlap (1997), Frick (1995, in press), and Robinson and Levin (1997) agreed that a measure of the size of an effect is indeed important in providing results to a reader.

We agree that it is important to provide an index of not only the statistical significance, but a measure of its magnitude. Robinson and Levin (1997) took the issue one step further and advocated for the use of adjectives such as *strong/large, moderate/medium*, etc. to refer to the effect size and to supply information concerning p values. However, some authors lead us to believe that they feel it may be necessary only to provide an index of practical significance and that it is unnecessary to provide statistical significance information. For example, it could be concluded from the writings of Carver (1978, 1993) and Shaver (1993) that they would like to abandon the use of statistical significance testing results. Although Cohen (1990, 1994) did not call for the outright abandonment of statistical significance testing, he did assert that you can attach a p-value to an effect size, but "it is far more informative to provide a confidence interval" (Cohen, 1990, p. 1310). Levin, in his 1993 article and in an article coauthored with Robinson (1997), argued against the idea of a single indicator of significance. Using hypothetical examples where the number of subjects in an experiment equals two, the authors provide evidence that practical significance, while noteworthy, does not provide evidence that the results gained were not gained by chance.

It is therefore the authors' opinion that it would be prudent to include both statistical significance and estimates of practical significance (not forgetting other important information such as evidence of replicability) within a research study. As Thompson (in press) discussed, any work undertaken in the social sciences will be based on subjective as well as objective criteria. The importance of subjective decision-making, as well as the idea that social science is imprecise and based on human judgment as well as objective criteria, helps to provide common benchmarks of quality. Subjectively choosing alpha levels (and in agreement with many researchers this does not necessarily denote a .05 or .01 level), power levels, and adjectives such as *large effects* for practical significance (cf. Cohen's [1988] treatise on power analysis, or Robinson and Levin's [1997] criteria for effect size estimates) are part of establishing common benchmarks or creating objective criteria. Robinson and Levin (1997) expressed the relationship between two types of significance quite succinctly: "First convince us that a finding is *not due to chance*, and only then, assess how *impressive* it is" (p. 23, italics in original).

Result Replicability

Carver (1978) was quick to identify that neither significance testing nor effect sizes typically inform the researcher regarding the likelihood that results will be replicated in future research. Schafer (1993), in response to the articles in *The Journal of Experimental Education*, felt that much of the criticism of significance testing was misfocused. Schafer concluded that readers of research should not mistakenly assume that statistical significance is an indication that the results may be replicated in the future; the issue of replication provides the impetus for the third recommendation provided by Thompson in both his 1989 *Measurement and Evaluation in Counseling and Development* article and 1996 AERA article.

According to Thompson (1996), "If science is the business of discovering replicable effects, because statistical significance tests do not evaluate result replicability, then researchers should use and report some strategies that *do* evaluate the replicability of their results" (p. 29, italics in original). Robinson and Levin (1997) were in total agreement with Thompson's recommendations of external result replicability. However, Robinson and Levin (1997) disagreed with Thompson when they concluded that internal replication analysis constitutes "an acceptable substitute for the genuine 'article'" (p. 26). Thompson (1997), in his rejoinder, recognized that external replication studies would be ideal in all situations, but concludes that many researchers do not have the stamina for external replication, and internal replicability analysis helps to determine where noteworthy results originate.

In terms of statistical significance testing, all of the arguments offered in the literature concerning replicability report that misconceptions about what statistical significance tells us are harmful to research. The authors of this paper agree, but once again note that misconceptions are a function of the researcher and not the test statistic. Replicability information offers important but somewhat different information concerning noteworthy results.

Importance of the Statistic as it Relates to Sample Size

According to Shaver (1993), a test of statistical significance "addresses only the simple question of whether a result is a likely occurrence under the null hypothesis with randomization and a sample of size *n*" (p. 301). Shaver's inclusion of "a sample of size *n*" indicates the importance of sample size in the H_0 decision-making process. As reported by Meehl (1967) and many authors since, with a large enough sample and reliable assessment, practically every association will be statistically significant. As noted previously, within Thompson's (1989) article a table was provided that showed the relationship between *n* and statistical significance when the effect size was kept constant. Two salient points applicable to this discussion were highlighted in Thompson's article: the first noted the relationship of *n* to statistical significance, providing a simulation that shows how, by varying *n* to create a large enough sample, a difference between two values can change a nonsignificant result into a statistically significant result. The second property of significance testing Thompson alluded to was an indication that "superficial understanding of significance testing has led to serious distortions, such as researchers interpreting significant results involving large effect sizes" (p. 2). Following this line of reasoning, Thompson (1993a) humorously noted that "tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and they are tired" (p. 363). Thus, as the sample size increases, the importance of significance testing is reduced. However, in small sample studies, significance testing can be useful, as it provides a level of protection from reporting random results by providing information about the chance of obtaining the sample statistics, given the sample size *n*, when the null hypothesis is exactly true in the population.

The Use of Language in Describing Results

Carver (1978, 1993), Cronbach (1975), Morrison and Henkel (1970), Robinson and Levin (1997), and Thompson (1987, 1989, 1993a, 1996, 1997) all stressed the need for the use of better language to describe significant results. As Schneider and Darcy (1984) and Thompson (1989) noted, significance is a function of at least seven interrelated features of a study where the size of the sample is the most influential characteristic. Thompson (1989) used an example of varying sample sizes with a fixed effect size to indicate how a small change in sample size affects the decision to reject, or fail to reject, H_0. The example helped to emphasize the cautionary nature that should be practiced in making judgements about the null hypothesis and raised the important issue of clarity in writing. These issues were the basis of Thompson's (1996) AERA article, where he called for the use of the term "statistically significant" when referring to the process of rejecting H_0 based on an alpha level. It was argued that through the use of specific terminology, the phrase "statistically significant" would not be confused with the common semantic meaning of *significant*.

In response, Robinson and Levin (1997) referred to Thompson's comments in the same light as Levin (1993) had done previously. While applauding Thompson for his "insightful analysis of the problem and the general spirit of each of his three article policy recommendations" (p. 21), Robinson and Levin were quick to counter with quips about "language police" and letting editors focus on content and substance and not on dotting the i's and crossing the t's. However, and interestingly, Robinson and Levin (1997) proceeded to concur with Thompson on the importance of language and continued their article with a call for researchers to use words that are more specific in nature. It is Robinson and Levin's (1997) recommendation that, instead of using the word statistically *significant*, researchers use statistically *nonchance* or statistically *real*, reflecting the test's intended meaning. The authors' rationale for changing the terminology reflects their wish to provide clear and precise information.

Thompson's (1997) rejoinder to the charges brought forth by Robinson and Levin (1997) was, fundamentally, to agree with their comments. In reference to the question of creating a "language police," Thompson admitted that

"I, too, find this aspect of my own recommendation troublesome" (p. 29). However, Thompson firmly believes the recommendations made in the AERA article should stand, citing the belief that "over the years I have reluctantly come to the conclusion that confusion over what statistical significance evaluates is sufficiently serious that an exception must be made in this case" (p. 29).

In respect to the concerns raised concerning the use of language, it is not the practice of significance testing that has created the statistical significance debate. Rather, the underlying problem lies with careless use of language and the incorrect assumptions made by less knowledgeable readers and practitioners of research. Cohen (1990) was quick to point out the rather sloppy use of language and statistical testing in the past, noting how one of the most grievous errors is the belief that the *p* value is the exact probability of the null hypothesis being true. Also, Cohen (1994) in his article; "The Earth is Round (*p* less than .05)" once again dealt with the ritual of null hypothesis significance testing and an almost mechanical dichotomous decision around a sacred $\alpha = .05$ criterion level. As before, Cohen (1994) referred to the misinterpretations that result from this type of testing (e.g., the belief that *p*-values are the probability that the null hypothesis is false). Cohen again suggested exploratory data analysis, graphical methods, and placing an emphasis on estimating effect sizes using confidence intervals. Once more, the basis for the argument against statistical significance testing falls on basic misconceptions of what the *p*-value statistic represents.

One of the strongest rationales for not using statistical significance values relies on misconceptions about the meaning of the *p*-value and the language used to describe its purpose. As Cortina and Dunlap (1997) noted, there are many cases where drawing conclusions based on *p* values are perfectly reasonable. In fact, as Cortina and Dunlap (1997), Frick (1995), Levin (1993), and Robinson and Levin (1997) pointed out, many of the criticisms of the *p* value are built on faulty premises, misleading examples, and incorrect assumptions concerning population parameters, null hypotheses, and their relationship to samples. For example, Cortina and Dunlap emphasized the incorrect use of logic (in particular the use of syllogisms and the Modus Tollens rule) in finding fault with significance testing, and Frick provides an interesting theoretical paper where he shows that in some circumstances, and based on certain assumptions, it *is* possible for the null hypothesis to be true.

It should be noted that several journals have adopted specific policies regarding the reporting of statistical results. The "Guidelines for Contributors" of the *Journal of Experimental Education* include the statement, "authors are *required* to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (Heldref Foundation, 1997, pp. 95-96, italics added). The *Educational and Psychological Measurement* "Guidelines for Authors" are even more emphatic. They state:

> We will go further [than mere encouragement]. Authors reporting statistical significance will be *required* to both report and interpret effect sizes. However, their effect sizes may be of various forms, including standardized differences, or uncorrected (e.g., r^2, R^2, eta^2) or corrected (e.g., adjusted R^2, omega^2) variance-accounted-for statistics. (Thompson, 1994, p. 845, italics in original)

At least one APA journal is also clear about this requirement. The following is from an editorial in the *Journal of Applied Psychology*.

> If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

For these journals, the reporting of effect size is required and the editors will consider statistical significance tests in their proper contexts. However, for most journals, the use of statistical and practical significance is determined by the views of the reviewers, and the editors and authors are subject to the decisions made by the reviewers they draw for their submissions.

The Recognition of the Importance of Other Types of Information

Other types of information are important when one considers statistical significance testing. The researcher should not ignore other information such as Type II errors, power analysis, and confidence intervals. While all of these statistical concepts are related, they provide different types of information that assist researchers in making decisions. There is an intricate relationship between power, sample size, effect size, and alpha (Cohen, 1988). Cohen recommended a power level of .80 for no other reason than that for which Fisher set an alpha level of .05 — it seemed a reasonable number to use. Cohen believed that the effect size should be set using theory, and the alpha level should be set using what degree of Type I error the researcher is willing to accept based on the type of experiment being conducted. In this scenario, *n* is the only value that may vary, and through the use of mathematical tables, is set at a particular value to be able to reach acceptable power, effect size, and alpha levels. Of course, in issues related to real-world examples, money is an issue and therefore sample sizes may be limited.

It is possible that researchers have to use small *n*'s because of the population they are studying (such as special education students). Cohen (1990) addresses the problems mentioned above by asking researchers to plan their research using the level of alpha risk they want to take, the size of the effect they wish to find, a calculated sample size, and the power they want. If one is unable to use a sample size of sufficient magnitude, one must compromise power, effect size, or as Cohen puts it, "even (heaven help us) increasing your alpha level" (p. 1310). This sentiment was shared by Schafer (1993) who—in reviewing the articles in the special issue of *The Journal of Experimental Education*—believed that researchers should set alpha levels, conduct power analysis, decide on the size of the sample, and design research studies that would increase effect sizes (e.g., through the careful addition of covariates in regression analysis or extending treatment interventions). It is necessary to balance sample size against power, and this automatically means that we do not fix one of them. It is also necessary to balance size and power against cost, which means that we do not arbitrarily fix sample size. All of the recommendations may be conducted prior to the data collection and therefore before the data analysis. The recommendations, in effect, provide evidence that methodological prowess may overcome some of the a posteriori problems researchers find.

Summary and Recommendations

We support other researchers who state that statistical significance testing must be accompanied by judgments of the event's practical significance and replicability. However, the likelihood of a chance occurrence of an event must not be ignored. We acknowledge the fact that the importance of significance testing is reduced as sample size increases. In large-sample experiments, particularly those involving multiple variables, the role of significance testing diminishes because even small, non-meaningful differences are often statistically significant. In small-sample studies where assumptions such as random sampling are practical, significance testing provides meaningful protection from random results. It is important to remember that statistical significance is only one criterion useful to inferential researchers. In addition to statistical significance, practical significance, and replicability, researchers must also consider Type II Errors and sample size. Furthermore, researchers should not ignore other techniques such as confidence intervals. While all of these statistical concepts are related, they provide different types of information that assist researchers in making decisions.

Our recommendations reflect a moderate mainstream approach. That is, we recommend that in situations where the assumptions are tenable, statistical significance testing still be applied. However, we recommend that the analyses always be accompanied by at least one measure of practical significance, such as effect size. The use of confidence intervals can be quite helpful in the interpretation of statistically significant or statistically nonsignificant results. Further, do not consider a hypothesis or theory "proven" even when both the statistical and practical significance have been established; the results have to be shown to be replicable. Even if it is not possible to establish external replicability for a specific study, internal approaches such as jackknife or bootstrap procedures are often feasible. Finally, please note that as sample sizes increase, the role of statistical significance becomes less important and the role of practical significance increases. This is because statistical significance can provide false comfort with results when sample sizes are large. This is especially true when the problem is multivariate and the large sample is representative of the target population. In these situations, effect size should weigh heavily in the interpretations.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Azar, B. (1997). *APA task force urges a harder look at data* [On-line]. Available: http://www.apa.org/monitor/mar97/stats.html

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61*(4), 287-292.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned so far. *American Psychologist, 45*(12), 1304-1312.

Cohen, J. (1994). The Earth is Round (*p* less than .05). *American Psychologist, 49*(12), 997-1003.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*(2), 161-172.

Cronbach, L. J. (1975). Beyond the two disciplines of psychology. *American Psychologist, 30*, 116-127.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition, 23*, 132-138.

Frick, R. W. (In press). Interpreting statistical testing: processes, not populations and random sampling. *Behavior Research Methods, Instruments, & Computers.*

Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory & Psychology, 1*, 375-382.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education, 61*(4), 317-333.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-59.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *The Journal of Experimental Education, 61*(4), 378-382.

Loftus, G. R., & Masson, M. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476-490.

Meehl P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Erlbaum.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3-5.

Robinson D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Schneider, A. L. & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation Review, 8*, 573-582.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education, 61*(4), 293-316.

Shea, C. (1996). Psychologists debate accuracy of "significance test." *Chronicle of Higher Education, 42*(49), A12, A16.

Snyder, P., & Lawson, S. (1993). Evaluating the results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education, 61*(4), 334-349.

Snyder, P. A., & Thompson, B. (in press). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly*.

Suen, H. K. (1992). Significance testing: Necessary but insufficient. *Topics in Early Childhood Special Education, 12(1)*, 66-81.

Task Force on Statistical Inference Initial Draft Report (1996). *Report to the Board of Scientific Affairs*. American Psychological Association [On-line]. Available: http://www.apa.org/science/tfsi.html.

Thompson, B. (1987, April). *The use (and misuse) of statistical significance testing. Some recommendations for improved editorial policy and practice*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868).

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22*, 2-5.

Thompson, B. (1993a). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*(4), 361-377.

Thompson, B. (Guest Ed.). (1993b). Statistical significance testing in contemporary practice [Special issue]. *The Journal of Experimental Education, 61*(4).

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29-32.

Thompson, B. (1998, January). *Why "encouraging" effect size reporting isn't working: The etiology of researcher resistance to changing practices*. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document ED Number forthcoming)

Thompson, B. (in press). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding multivariate statistics* (*Vol. 2*). Washington, DC: American Psychological Association.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the *Journal of Experimental Education. Journal Experimental Education, 66*, 75-83.

Thompson, B., & Snyder, P. A. (in press). Statistical significance testing and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development*.

Vacha-Haase, T., & Nilsson, J. E. (in press). Statistical significance reporting: Current trends and usages within *MECD*. *Measurement and Evaluation in Counseling and Development*.

Weitzman, R. A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports, 54*, 355-363.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 23-32

Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals

Larry G. Daniel
University of North Texas

Statistical significance tests (SSTs) have been the object of much controversy among social scientists. Proponents have hailed SSTs as an objective means for minimizing the likelihood that chance factors have contributed to research results; critics have both questioned the logic underlying SSTs and bemoaned the widespread misapplication and misinterpretation of the results of these tests. The present paper offers a framework for remedying some of the common problems associated with SSTs via modification of journal editorial policies. The controversy surrounding SSTs is overviewed, with attention given to both historical and more contemporary criticisms of bad practices associated with misuse of SSTs. Examples from the editorial policies of Educational and Psychological Measurement and several other journals that have established guidelines for reporting results of SSTs are overviewed, and suggestions are provided regarding additional ways that educational journals may address the problem.

Statistical significance testing has existed in some form for approximately 300 years (Huberty, 1993) and has served an important purpose in the advancement of inquiry in the social sciences. However, there has been much controversy over the misuse and misinterpretation of statistical significance testing (Daniel, 1992b). Pedhazur and Schmelkin (1991, p. 198) noted, "Probably few methodological issues have generated as much controversy among sociobehavioral scientists as the use of [statistical significance] tests." This controversy has been evident in social science literature for some time, and many of the articles and books exposing the problems with statistical significance have aroused remarkable interest within the field. In fact, at least two articles on the topic appeared in a list of works rated by the editorial board members of *Educational and Psychological Measurement* as most influential to the field of social science measurement (Thompson & Daniel, 1996b). Interestingly, the criticisms of statistical significance testing have been pronounced to the point that, when one reviews the literature, "it is more difficult to find specific arguments for significance tests than it is to find arguments decrying their use" (Henkel, 1976, p. 87); nevertheless, Harlow, Mulaik, and Steiger (1997), in a new book on the controversy, present chapters on both sides of the issue. This volume, titled *What if There Were No Significance Tests?*, is highly recommended to those interested in the topic, as is a thoughtful critique of the volume by Thompson (1998).

Thompson (1989b) noted that researchers are increasingly becoming aware of the problem of over-reliance on statistical significance tests (referred to herein as "SSTs"). However, despite the influence of the many works critical of practices associated with SSTs, many of the problems raised by the critics are still prevalent. Researchers have inappropriately utilized statistical significance as a means for illustrating the importance of their findings and have attributed to statistical significance testing qualities it does not possess. Reflecting on this problem, one psychological researcher observed, "the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; . . . a great deal of mischief has been associated with its use" (Bakan, 1966, p. 423).

Because SSTs have been so frequently misapplied, some reflective researchers (e.g., Carver, 1978; Meehl, 1978; Schmidt, 1996; Shulman, 1970) have recommended that SSTs be completely abandoned as a method for evaluating statistical results. In fact, Carver (1993) not only recommended abandoning statistical significance testing, but referred to it as a "corrupt form of the scientific method" (p. 288). In 1996, the American Psychological Association (APA) appointed its Task Force on Statistical Inference, which considered among other actions recommending less or even no use of statistical significance testing within APA journals (Azar, 1997; Shea, 1996). Interestingly, in its draft report, the Task Force (Board of Scientific Affairs, 1996) noted that it "does not support any action that could be interpreted as banning the use of null hypothesis significance testing" (p. 1). Furthermore, SSTs still have support from a number

---

Larry G. Daniel is a professor of education at the University of North Texas. The author is indebted to five anonymous reviewers whose comments were instrumental in improving the quality of this paper. Address correspondence to Larry G. Daniel, University of North Texas, Denton, TX 76203 or by e-mail to daniel@tac.coe.unt.edu.

of reflective researchers who acknowledge their limitations, but also see the value of the tests when appropriately applied. For example, Mohr (1990) reasoned, "one cannot be a slave to significance tests. But as a first approximation to what is going on in a mass of data, it is difficult to beat this particular metric for communication and versatility" (p. 74). In similar fashion, Huberty (1987) maintained, "there is nothing wrong with statistical tests themselves! When used as guides and indicators, as opposed to a means of arriving at definitive answers, they are okay" (p. 7).

"Statistical Significance" Versus "Importance"

A major controversy in the interpretation of SSTs has been "the ingenuous assumption that a statistically significant result is necessarily a noteworthy result" (Daniel, 1997, p. 106). Thoughtful social scientists (e.g., Berkson, 1942; Chow, 1988; Gold, 1969; Shaver, 1993; Winch & Campbell, 1969) have long recognized this problem. For example, even as early as 1931, Tyler had already begun to recognize a trend toward the misinterpretation of statistical significance:

> The interpretations which have commonly been drawn from recent studies indicate clearly that we are prone to conceive of statistical significance as equivalent to social significance. These two terms are essentially different and ought not to be confused. . . . Differences which are statistically significant are not always socially important. The corollary is also true: differences which are not shown to be statistically significant may nevertheless be socially significant. (pp. 115-117)

A decade later, Berkson (1942) remarked, "statistics, as it is taught at present in the dominant school, consists almost entirely of tests of significance" (p. 325). Likewise, by 1951, Yates observed, "scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and this is the end of it" (p. 33). Similarly, Kish (1959) bemoaned the fact that too much of the research he had seen was presented "at the primitive level" (p. 338). Twenty years later, Kerlinger (1979) recognized that the problem still existed:

> statistical significance says little or nothing about the magnitude of a difference or of a relation. With a large number of subjects . . . tests of significance show statistical significance even when a difference between means is quite small, perhaps trivial, or a correlation coefficient is very small and trivial. . . . To use statistics adequately, one must understand the principles involved and be able to judge whether obtained results are statistically significant *and* whether they are meaningful in the particular research context. (pp. 318-319, emphasis in original)

Contemporary scholars continue to recognize the existence of this problem. For instance, Thompson (1996) and Pedhazur and Schmelkin (1991) credit the continuance of the misperception, in part, to the tendency of researchers to utilize and journals to publish manuscripts containing the term "significant" rather than "statistically significant"; thus, it becomes "common practice to drop the word 'statistical,' and speak instead of 'significant differences,' 'significant correlations,' and the like" (Pedhazur & Schmelkin, 1991, p. 202). Similarly, Schafer (1993) noted, "I hope most researchers understand that *significant* (statistically) and *important* are two different things. Surely the term *significant* was ill chosen" (p. 387, emphasis in original). Moreover, Meehl (1997) recently characterized the use of the term "significant" as being "cancerous" and "misleading" (p. 421) and advocated that researchers interpret their results in terms of confidence intervals rather than p values.

SSTs and Sample Size

Most tests of statistical significance utilize some test statistic (e.g., F, t, chi-square) with a known distribution. An SST is simply a comparison of the value for a particular test statistic based on results of a given analysis with the values that are "typical" for the given test statistic. The computational methods utilized in generating these test statistics yield larger values as sample size is increased, given a fixed effect size. In other words, for a given statistical effect, a large sample is more likely to guarantee the researcher a statistically significant result than a small sample is. For example, suppose a researcher was investigating the correlation between scores for a given sample on two tests. Hypothesizing that the tests would be correlated, the researcher posited the null hypothesis that r would be equal to zero. As illustrated in Table 1, with an extremely small sample, even a rather appreciable r-value would not be statistically significant ($p < .05$). With a sample of only 10 persons, for example, an r as large as .6, indicating a moderate to large statistical effect, would not be statistically significant; by contrast, a negligible statistical effect of less than 1% ($r^2 = .008$) would be statistically significant with a sample size of 500!

Table 1
Critical Values of r for Rejecting the Null Hypothesis ($r = 0$) at the .05 Level Given Sample Size n

n	r
3	.997
5	.878
10	.632
20	.444
50	.276
100	.196
500	.088
1,000	.062
5,000	.0278
10,000	.0196

Note: Values are taken from Table 13 in Pearson and Hartley (1962).

As a second example, suppose a researcher is conducting an educational experiment in which students are randomly assigned to two different instructional settings and are then evaluated on an outcome achievement measure. This researcher might utilize an analysis of variance test to evaluate the result of the experiment. Prior to conducting the test (and the experiment), the researcher would propose a null hypothesis of no difference between persons in varied experimental conditions and then compute an F statistic by which the null hypothesis may be evaluated. F is an intuitively-simple ratio statistic based on the quotient of the mean square for the effect(s) divided by the mean square for the error term. Since mean squares are the result of dividing the sum of squares for each effect by its degrees of freedom, the mean square for the error term will get smaller as the sample size is increased and will, in turn, serve as a smaller divisor for the mean square for the effect, yielding a larger value for the F statistic. In the present example (a two-group, one-way ANOVA), a sample of 302 would be five times as likely to yield a statistically significant result as a sample of 62 simply due to a larger number of error degrees of freedom (300 versus 60). In fact, with a sample as large as 302, even inordinately trivial differences between the two groups could be statistically significant considering that the p value associated with a large F will be small.

As these examples illustrate, an SST is largely a test of whether or not the sample is large, a fact that the researcher knows even before the experiment takes place. Put simply, "Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects" (Thompson, 1992, p. 436). Some 60 years ago, Berkson (1938, pp. 526-527) exposed this circuitous logic based on his own observation of statistical significance values associated with chi-square tests with approximately 200,000 subjects:

> an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small . . . and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square P will be small if the sample has a sufficiently large number of observations it. . . . If, then, we know in advance the P that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all!

Misinterpretation of the Meaning of "Statistically Significant"

An analysis of past and current social science literature will yield evidence of at least six common misperceptions about the meaning of "statistically significant." The first of these, that "statistically significant" means "important," has already been addressed herein. Five additional misperceptions will also be discussed briefly: (a) the misperception that statistical significance informs the researcher as to the likelihood that a given result will be replicable ("the replicability fantasy" – Carver, 1978); (b) the misperception that statistical significance informs the researcher as to the likelihood that results were due to chance (or, as Carver [1978, p. 383] termed it, "the odds-against-chance fantasy"); (c) the misperception that a statistically significant result indicates the likelihood that the sample employed is representative of the population; (d) the misperception that statistical significance is the best way to evaluate statistical results; and (e) the misperception that statistically significant reliability and validity coefficients based on scores on a test administered to a given sample imply that the same test will yield valid or reliable scores with a different sample.

SSTs and replicability. Despite misperceptions to the contrary, the logic of statistical significance testing is NOT an appropriate means for assessing result replicability (Carver, 1978; Thompson, 1993a). Statistical significance simply indicates the probability that the null hypothesis is true in the population. However, Thompson (1993b) provides discussion of procedures that may provide an estimate of replicability. These procedures (cross validation, jackknife methods, and bootstrap methods) all involve sample-splitting logics and allow for the computation of statistical estimators across multiple configurations of the same sample in a single study. Even though these methods are biased to some degree (a single sample is utilized in each of the procedures), they represent the next best alternative to conducting a replication of the given study (Daniel, 1992a). Ferrell (1992) demonstrated how results from a single multiple regression analysis can be cross validated by randomly splitting the original sample and predicting dependent variable scores for each half of the sample using the opposite group's weights. Daniel (1989) and Tucker and Daniel (1992) used a similar logic in their analyses of the generalizability of results with the sophisticated "jackknife" procedure. Similar heuristic presentations of the computer-intensive "bootstrap" logic are also available in the extant literature (e.g., Daniel, 1992a).

SSTs and odds against chance. This common misperception is based on the naive perception that statistical significance measures the degree to which results of a given SST occur by chance. By definition, an SST tests the probability that a null hypothesis (i.e., a hypothesis positing no relationship between variables or no difference between groups) is true in a given population based on the results of a sample of size *n* from that population. Consequently, "a test of significance provides the *probability of a result occurring by chance in the long run under the null hypothesis* with random sampling and sample size *n*; it provides *no basis for a conclusion about the probability that a given result is attributable to chance*" (Shaver, 1993, p. 300, emphasis added). For example, if a correlation coefficient r of .40 obtained between scores on Test X and Test Y for a sample of 100 fifth graders is statistically significant at the 5% ($\alpha = .05$) level, one would appropriately conclude that there is a 95% likelihood that the correlation between the tests in the population is not zero assuming that the sample employed is representative of the population. However, it would be *inappropriate* to conclude (a) that there is a 95% likelihood that the correlation is .40 in the population or (b) that there is only a 5% likelihood that the result of that particular statistical significance test is due to chance. This fallacy was exposed by Carver (1978):

> the p value is the probability of getting the research results when it is first assumed that it is actually true that chance caused the results. It is therefore impossible for the p value to be the probability that chance caused the mean difference between two research groups since (a) the p value was calculated by assuming that the probability was 1.00 that chance did cause the mean difference, and (b) the p value is used to decide whether to accept or reject the idea that probability is 1.00 that chance caused the mean difference. (p. 383)

SSTs and sampling. This misperception states that the purpose of statistical significance testing is to determine the degree to which the sample represents the population. Representativeness of the sample cannot be evaluated with an SST; the only way to estimate if a sample is representative is to carefully select the sample. In fact, the statistical significance test is better conceptualized as answering the question, "*If* the sample represents the population, how likely is the obtained result?"

SSTs and evaluation of results. This misperception, which states that the best (or correct) way to evaluate the statistical results is to consult the statistical significance test, often accompanies the "importance" misperception but actually may go a step beyond the importance misperception in its corruptness. The importance misperception, as previously noted, simply places emphasis on the wrong thing. For example, the researcher might present a table of correlations, but in interpreting and discussing the results, only discuss whether or not each test yielded a statistically significant result, making momentous claims for statistically significant correlations no matter how small and ignoring statistically nonsignificant values no matter how large. In this case, the knowledgeable reader could still look at the correlations and draw more appropriate conclusions based on the magnitude of the r values. However, if the researcher were motivated by the "result evaluation" misperception, he or she might go so far as to fail to report the actual correlation values, stating only that certain relationships were statistically significant. Likewise, in the case of an analysis of variance, this researcher might simply report the F statistic and its p value without providing a breakdown of the dependent variable sum of squares from which an estimate of effect size could be determined. Thompson (1989a, 1994) discussed several suggestions for improvement of these practices, including the reporting of (a) effect sizes for all parametric analyses and (b) "what if" analyses "indicating at what different sample size a given fixed effect would become statistically significant or would have no longer been statistically significant" (1994, p. 845). In regard to (b), Morse (1998) has designed a PC-compatible computer program for assessing the sensitivity of results to sample size. Moreover, in the cases in which statistically nonsignificant results are obtained, researchers should consider conducting statistical power analyses (Cohen, 1988).

SSTs and test score characteristics. Validity and reliability are characteristics of test *scores* or test *data*.

However, contemporary scholarly language (e.g., "the test is reliable," "the test is valid") often erroneously implies that validity and reliability are characteristics of tests themselves. This fallacious use of language is sometimes accompanied by another fallacy related to statistical significance testing, namely, the use of null hypothesis SSTs of reliability or validity coefficients. Statistical tests of these coefficients are nonsensical. As Witta and Daniel (1998) noted:

> In the case of a reliability coefficient, these statistical significance tests evaluate the null hypothesis that a set of scores is totally unreliable, a hypothesis that is meaningless considering that large reliability or validity coefficients may often be statistically significant even when based on extremely small samples (Thompson, 1994) whereas minute reliability or validity coefficients will eventually become statistically significant if the sample size is increased to a given level (Huck & Cormier, 1996). Further, considering that reliability and validity coefficients are sample specific, statistical significance tests do not offer any promise of the generalizability of these coefficients to other samples. (pp. 4-5)

Journal Policies and Statistical Significance

As most educational researchers are aware, social science journals have for years had a bias towards accepting manuscripts documenting statistically significant findings and rejecting those with statistically nonsignificant findings. One editor even went so far as to boast that he had made it a practice to avoid accepting for publication results that were statistically significant at the .05 level, desiring instead that results reached at least the .01 level (Melton, 1962). Because of this editorial bias, many researchers (e.g., Mahoney, 1976) have paid homage to SSTs in public while realizing their limitations in private. As one observer noted a generation ago, "Too, often . . . even wise and ingenious investigators, for varieties of reasons, not the least of which are the editorial policies of our major psychological journals, . . . tend to credit the test of significance with properties it does not have" (Bakan, 1966, p. 423).

According to many researchers (e.g., Neuliep, 1991; Shaver, 1993), this bias against studies that do not report statistical significance or that present results that did not meet the critical alpha level still exists. Shaver (1993) eloquently summarized this problem:

> Publication is crucial to success in the academic world. Researchers shape their studies, as well as the manuscripts reporting the research, according to accepted ways of thinking about analysis and interpretation and to fit their perceptions of what is publishable. To break from the mold might be courageous, but, at least for the untenured faculty member with some commitment to self-interest, foolish. (p. 310)

Because this bias is so prevalent, it is not uncommon to find examples in the literature of studies that report results that are statistically nonsignificant with the disclaimer that the results "approached significance." Thompson (1993a) reported a somewhat humorous, though poignant, response by one journal editor to this type of statement: "How do you know your results were not working very hard to *avoid* being statistically significant?" (p. 285, emphasis in original).

Likewise, results that are statistically significant at a conservative alpha level (e.g, .001), are with some frequency referred to as "highly significant," perhaps with the authors' intent being to make a more favorable impression on some journal editors and readers than they could make by simply saying that the result was statistically significant, period. This practice, along with the even more widespread affinity for placing more and more zeroes to the right of the decimal in an attempt to make a calculated p appear more noteworthy, has absolutely nothing to do with the practical significance of the result. The latter practice has often been the focus of tongue-in-cheek comments. For example, Popham (1993) noted, "Some evaluators report their probabilities so that they look like the scoreboard for a no-hit baseball game (e.g., $p < .000000001$)" (p. 266); Campbell (1982) quipped, "It is almost impossible to drag authors away from their p values, and the more zeroes after the decimal point, the harder people cling to them" (p. 698); and McDonald (1985), referring to the tendency of authors to place varying numbers of stars after statistical results reported in tabular form as a means for displaying differing levels of statistical significance, bantered that the practice resembled "grading of hotels in guidebooks" (p. 20).

If improvements are to be made in the interpretation and use of SSTs, professional journals (Rozeboom, 1960), and, more particularly, their editors will no doubt have to assume a leadership role in the effort. As Shaver (1993) articulated it, "As gatekeepers to the publishing realm, journal editors have tremendous power. . .[and perhaps should] become crusaders for an agnostic, if not atheistic, approach to tests of statistical significance" (pp. 310-311). Hence, Carver (1978, 1993) and Kupfersmid (1988) suggested that journal editors are the most likely candidates to promote an end to the misuse and misinterpretation of SSTs.

Considering this, it is encouraging to note that at least some journals have begun to adopt policies relative to statistical significance testing that address some of the problems discussed here. For several years, *Measurement and Evaluation in Counseling and Development* (1992, p. 143) has included three specific (and appropriate) author guidelines related to statistical significance testing, including the encouragement for authors to (a) index results of SSTs to sample size, (b) provide readers with effect size estimates as well as SSTs, and (c) provide power estimates of protection against Type II error when statistically nonsignificant results are obtained.

Educational and Psychological Measurement (*EPM*) has developed a similar set of editorial policies (Thompson, 1994) which are presently in their fourth year of implementation. These guidelines do not for the most part ban the use of SSTs from being included in authors' manuscripts, but rather request that authors report other information along with the SST results. Specifically, these editorial guidelines include the following:

1. Requirement that authors use "statistically significant" and not merely "significant" in discussing results.
2. Requirement that tests of statistical significance generally NOT accompany validity and reliability coefficients (Daniel & Witta, 1997; Huck & Cormier, 1996; Witta & Daniel, 1998). This is the one scenario in which SSTs are expressly forbidden according to *EPM* editorial policy.
3. Requirement that all statistical significance tests be accompanied by effect size estimates.
4. Suggestion that authors may wish to report the "what if" analyses alluded to earlier. These analyses should indicate "at what different sample size a given fixed effect would become statistically significant or would have no longer been statistically significant" (Thompson, 1994, p. 845).
5. Suggestion that authors report external replicability analyses via use of data from multiple samples or else internal replicability analyses via use of cross-validation, jackknife, or bootstrap procedures.

A number of efforts have been utilized by the *EPM* editors to help both authors and reviewers become familiar with these guidelines. For the first two years that these guidelines were in force, copies of the guidelines editorial (Thompson, 1994) were sent to every author along with the manuscript acceptance letter. Although copies are no longer sent to authors, the current manuscript acknowledgment letter includes a reference to this and two other author guidelines editorials the journal has published (Thompson, 1995; Thompson & Daniel, 1996a), and it directs authors to refer to the several editorials to determine if their manuscripts meet editorial policy. More recently, the several editorials have been made available via the Internet at Web address: "http://acs.tamu.edu/~bbt6147/".

In addition to this widescale distribution policy, the guidelines are referenced on each review form (see Appendix) sent to the masked reviewers. As a part of the review process, reviewers must determine if manuscripts contain material that is in violation of the editorial policies relative to statistical significance testing and several other methodological issues. To assure that reviewers will take this responsibility seriously, several questions relative to the guidelines editorials are included on the review form and must be answered by the reviewers. No manuscripts are accepted for publication by either of the two current editors if they violate these policies, although these violations do not necessarily call for outright rejection of the first draft of a manuscript. It is the hope of the editors that this comprehensive policy will over time make a serious impact on *EPM* authors' and readers' ideas about correct practice in reporting the results of SSTs.

More recently, two additional journals have adopted editorial policies that are likely to prompt additional scrutiny of the reporting and interpretation of SSTs. The current author guidelines of the *Journal of Experimental Education* (Heldref Foundation, 1997) indicate that "authors are *required* to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (pp. 95-96, emphasis added). Further, the editor of one of the APA journals, *Journal of Applied Psychology*, recently stated:

> If an author decides not to report an effect size estimate along with the outcome of a [statistical] significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

Recommendations for Journal Editors

As the previous discussion has illustrated, there is a trend among social science journal editors to either reject or demand revision of manuscripts in which authors employ loose language relative to their interpretations of SSTs or else overinterpret the results of these tests; however, more movement of the field toward this trend is needed. Pursuant to the continued movement toward this trend, the following ten recommendations are offered to journal editors and scholars at large as a means for encouraging better practices in educational journals and other social science journals.

1. *Implement editor and reviewer selection policies.* First, following the suggestions of Carver (1978, 1993) and Shaver (1993), it would be wise for professional associations and publishers who hire/appoint editors for their publications to require potential editors to submit statements relative to their positions on statistical significance testing. Journal editors might also require a similar statement from persons who are being considered as members of editorial review boards.
2. *Develop guidelines governing SSTs.* Each editor should adopt a set of editorial guidelines that will promote correct practice relative to the use of SSTs. The *Measurement and Evaluation in Counseling and Development* and *Educational and Psychological Measurement* guidelines referenced in this paper could serve as a model for policies developed for other journals.
3. *Develop a means for making the policies known to all involved.* Editors should implement a mechanism whereby authors and reviewers will be likely to remember and reflect upon the policies. The procedures mentioned previously that are currently utilized by the editors of *Educational and Psychological Measurement* might serve as a model that could be adapted to the needs of a given journal.
4. *Enforce current APA guidelines for reporting SSTs.* Considering that most journals in education and psychology utilize APA publication guidelines, editors could simply make it a requirement that the guidelines for reporting results of SSTs included in the fourth edition *Publication Manual of the American Psychological Association* (APA, 1994, pp. 17-18) be followed. Although the third edition *Publication Manual* was criticized for using statistical significance reporting examples that were flawed (Pedhazur & Schmelkin, 1991; Shaver, 1993), the fourth edition includes appropriate examples as well as suggestions encouraging authors to report effect size estimates.
5. *Require authors to use "statistically" before "significant."* Despite the fact that some journal editors will be resistant to the suggestion (see, for example, Levin's [1993; Robinson & Levin, 1997] criticism that such a practice smacks of policing of language), requiring authors to routinely use the term "statistically significant" rather than simply "significant" (cf. Carver, 1993; Cohen, 1994; Daniel, 1988; Shaver, 1993; Thompson, 1996) when referring to research findings will do much to minimize the "statistical significance as importance" problem and to make it clear where the author intends to make claims about the "practical significance" (Kirk, 1996) of the results.
6. *Require effect size reporting.* Editors should require that effect size estimates be reported for all quantitative analyses. These are strongly *suggested* by APA (1994); however, Thompson (1996, p. 29, emphasis in original) advocated that other professional associations that publish professional journals "venture beyond APA, and *require* such reports in all quantitative analyses."
7. *Encourage or require replicability and "what if" analyses.* As previously discussed, replicability analyses provide reasonable evidence to support (or disconfirm) the generalizability of the findings, something that SSTs do NOT do (Shaver, 1993; Thompson, 1994). "What if" analyses, if used regularly, will build in readers and authors a sense of always considering the sample size when conducting SSTs, and thereby considering the problems inherent in particular to cases involving rather larger or rather small samples.
8. *Require authors to avoid using SSTs where they are not appropriate.* For example, as previously noted, *EPM* does not allow manuscripts to be published if SSTs accompany certain validity or reliability coefficients.
9. *Encourage or require that power analyses or replicability analyses accompany statistically nonsignificant results.* These analyses allow for the researcher to address power considerations or to determine if a result with a small sample has evidence of stability in cases in which an SST indicates a statistically nonsignificant result.
10. *Utilize careful copyediting procedures.* Careful copyediting procedures will serve to assure that very little sloppy language relative to SSTs will end up in published manuscripts. In addition to the suggestions mentioned above, editors will want to make sure language such as "highly significant" and "approaching significance" is edited out of the final copies of accepted manuscripts.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington: Author.

Azar, B. (1997). APA task force urges a harder look at data. *APA Monitor, 28*(3), 26.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

Board of Scientific Affairs. (1996). *Task Force on Statistical Inference initial report (DRAFT)* [On-line]. Available: http://www.apa.org/science/tsfi/html

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67*, 691-700.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 70*, 426-443.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Daniel, L. G. (1988). [Review of *Conducting educational research* (3rd ed.)]. *Educational and Psychological Measurement, 48*, 848-851.

Daniel, L. G. (1989, January). *Use of the jackknife statistic to establish the external validity of discriminant analysis results*. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 305 382)

Daniel, L. G. (1992a, April). *Bootstrap methods in the principal components case*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 346 135)

Daniel, L. G. (1992b, November). *Perceptions of the quality of educational research throughout the twentieth century: A comprehensive review of the literature*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Daniel, L. G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. *Journal of Experimental Education, 65*, 101-112.

Daniel, L. G., & Witta, E. L. (1997, March). *Implications for teaching graduate students correct terminology for discussing validity and reliability based on a content analysis of three social science measurement journals*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 408 853)

Ferrell, C. M. (1992, February). *Statistical significance, sample splitting and generalizability of results*. Paper presented at the annual meeting of the Southwest Educational Research Association. (ERIC Document Reproduction Service No. ED 343 935)

Gold, D. (1969). Statistical tests and substantive significance. *American Sociologist, 4*, 42-46.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

Henkel, C. G. (1976). *Tests of significance*. Newbury Park, CA: Sage.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher, 16*(8), 4-9.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317-333.

Huck, S. W., & Cormier, W. G. (1996). *Reading statistics and research* (2nd ed.). New York: HarperCollins.

Kerlinger, F. N. (1979). *Behavioral research: A conceptual approach*. New York: Holt, Rinehart and Winston.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 5*, 746-759.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review, 24*, 328-338.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635-642.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*, 378-382.

Mahoney, M. J. (1976). *Scientist as subject: The psychological imperative*. Cambridge, MA: Ballinger.

McDonald, R. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

Measurement and Evaluation in Counseling and Development. (1992). Guidelines for authors. *Measurement and Evaluation in Counseling and Development, 25*, 143.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-426). Mahwah, NJ: Erlbaum.

Melton, A. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Mohr, L. B. (1990). *Understanding significance testing.* Newbury Park, CA: Sage.

Morse, D. T. (1998). MINSIZE: A computer program for obtaining minimum sample size as an indicator of effect size. *Educational and Psychological Measurement, 58*, 142-153.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3-5.

Neuliep, J. W. (Ed.). (1991). *Replication in the social sciences.* Newbury Park, CA: Sage.

Pearson, E. S., & Hartley, H. O. (Eds.). (1962). *Biometrika tables for statisticians* (2nd ed.). Cambridge, MA: Cambridge University Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Popham, W. J. (1993). *Educational evaluation* (3rd ed.). Boston, MA: Allyn and Bacon.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rozeboom, W. M. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance, *Journal of Experimental Education, 61*, 383-387.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*(2), 115-129.

Shaver, J. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*, 293-316

Shea, C. (1996). Psychologists debate accuracy of "significance" test. *Chronicle of Higher Education, 42*(9), A12,A19.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40*, 371-393.

Thompson, B. (1989a). Asking "what if" questions about significance tests. *Measurement and Evaluation in Counseling and Development, 22*, 66-67.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22*, 2-6.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B. (1993a). Foreword. *Journal of Experimental Education, 61*, 285-286.

Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998). [Review of *What if there were no significance tests?*]. *Educational and Psychological Measurement, 58*, 334-346.

Thompson, B., & Daniel, L. G. (1996a). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Thompson, B., & Daniel, L. G. (1996b). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement, 56*, 741-745.

Tucker, M. L., & Daniel, L. G. (1992, January). *Investigating result stability of canonical function equations with the jackknife technique.* Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 343 914)

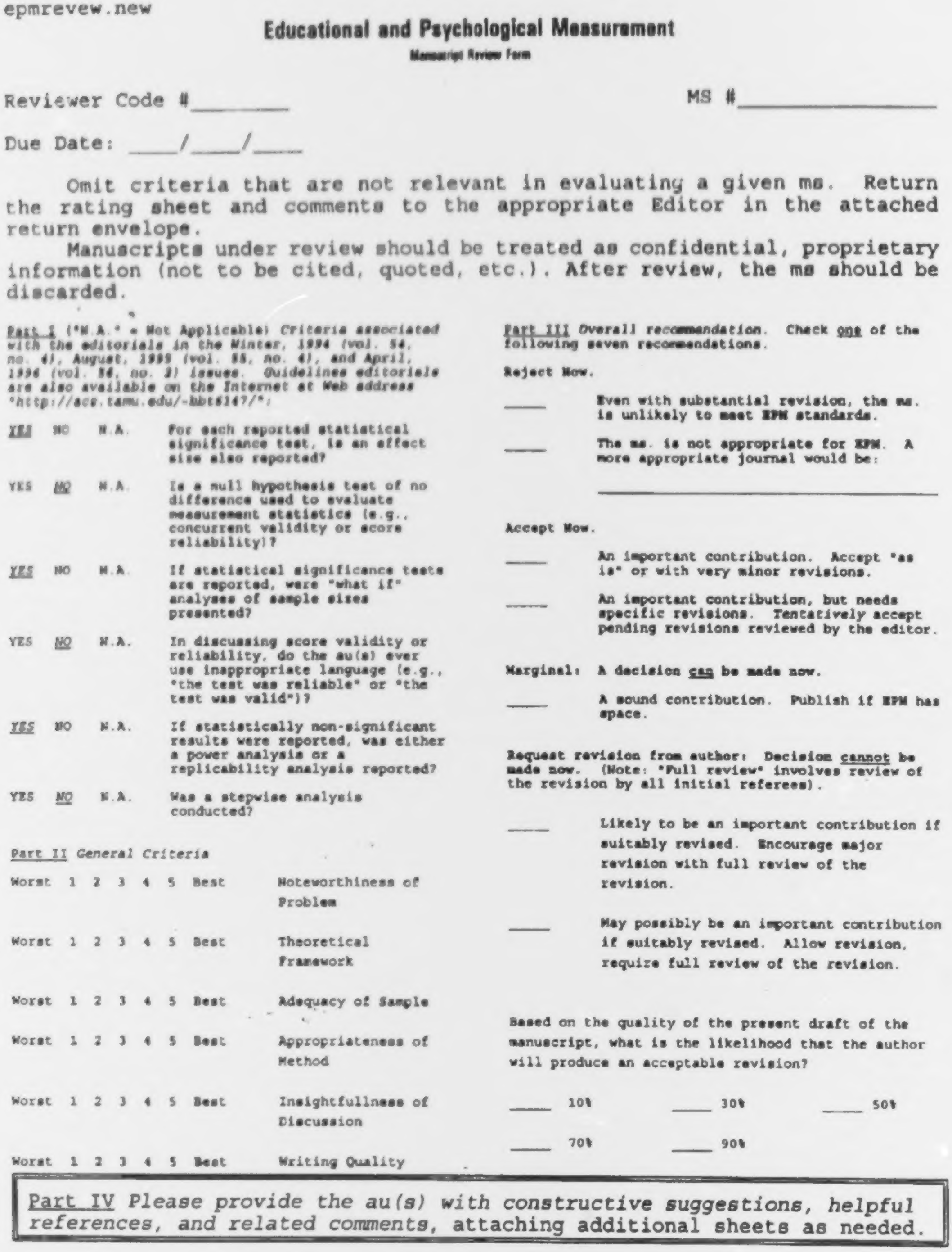
Tyler, R. W. (1931). What is statistical significance? *Educational Research Bulletin, 10*, 115-118,142.

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist, 4*, 140-143.

Witta, E. L., & Daniel, L. G. (1998, April). *The reliability and validity of test scores: Are editorial policy changes reflected in journal articles?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.

APPENDIX
EPM MANUSCRIPT REVIEW FORM



epmrevew.new

Educational and Psychological Measurement

Manuscript Review Form

Reviewer Code #________ MS #________________

Due Date: ____/____/____

Omit criteria that are not relevant in evaluating a given ms. Return the rating sheet and comments to the appropriate Editor in the attached return envelope.

Manuscripts under review should be treated as confidential, proprietary information (not to be cited, quoted, etc.). After review, the ms should be discarded.

Part I ("N.A." = Not Applicable) *Criteria associated with the editorials in the Winter, 1994 (vol. 54, no. 4), August, 1995 (vol. 55, no. 4), and April, 1996 (vol. 56, no. 2) issues. Guidelines editorials are also available on the Internet at Web address "http://acs.tamu.edu/~bbt6147/":*

[✓ YES]	NO	N.A.	For each reported statistical significance test, is an effect size also reported?
YES	[✓ NO]	N.A.	Is a null hypothesis test of no difference used to evaluate measurement statistics (e.g., concurrent validity or score reliability)?
[✓ YES]	NO	N.A.	If statistical significance tests are reported, were "what if" analyses of sample sizes presented?
YES	[✓ NO]	N.A.	In discussing score validity or reliability, do the au(s) ever use inappropriate language (e.g., "the test was reliable" or "the test was valid")?
[✓ YES]	NO	N.A.	If statistically non-significant results were reported, was either a power analysis or a replicability analysis reported?
YES	[✓ NO]	N.A.	Was a stepwise analysis conducted?

Part II *General Criteria*

Worst 1 2 3 4 5 Best	Noteworthiness of Problem
Worst 1 2 3 4 5 Best	Theoretical Framework
Worst 1 2 3 4 5 Best	Adequacy of Sample
Worst 1 2 3 4 5 Best	Appropriateness of Method
Worst 1 2 3 4 5 Best	Insightfullness of Discussion
Worst 1 2 3 4 5 Best	Writing Quality

Part III *Overall recommendation.* Check one of the following seven recommendations.

Reject Now.

_____ Even with substantial revision, the ms. is unlikely to meet EPM standards.

_____ The ms. is not appropriate for EPM. A more appropriate journal would be:

____

Accept Now.

_____ An important contribution. Accept "as is" or with very minor revisions.

_____ An important contribution, but needs specific revisions. Tentatively accept pending revisions reviewed by the editor.

Marginal: A decision can be made now.

_____ A sound contribution. Publish if EPM has space.

Request revision from author: Decision cannot be made now. (Note: "Full review" involves review of the revision by all initial referees).

_____ Likely to be an important contribution if suitably revised. Encourage major revision with full review of the revision.

_____ May possibly be an important contribution if suitably revised. Allow revision, require full review of the revision.

Based on the quality of the present draft of the manuscript, what is the likelihood that the author will produce an acceptable revision?

_____ 10% _____ 30% _____ 50%

_____ 70% _____ 90%

Part IV *Please provide the au(s) with constructive suggestions, helpful references, and related comments,* attaching additional sheets as needed.

BEST COPY AVAILABLE

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 33-38

Statistical Significance and Effect Size Reporting: Portrait of a Possible Future

Bruce Thompson
Texas A&M University and Baylor College of Medicine

The present paper comments on the matters raised regarding statistical significance tests by three sets of authors in this issue. These articles are placed within the context of contemporary literature. Next, additional empirical evidence is cited showing that the APA publication manual's "encouraging" effect size reporting has had no appreciable effect. Editorial policy will be required to affect change, and some model policies are quoted. Science will move forward to the extent that both effect size and replicability evidence of one or more sorts are finally seriously considered within our inquiry.

I appreciate the opportunity to comment on matters raised by Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998) as regards statistical significance tests. Theme issues of journals such as the present one (see also Thompson (1993)) allow various perspectives to be articulated and help slowly but inexorably move the field toward improved practices. Of course, an important recent book (Harlow, Mulaik, & Steiger, 1997) also presents diverse perspectives regarding these continuing controversies (for reviews see Levin (1998) and Thompson (1998c)).

At the outset perhaps I should acknowledge possible conflicts of interest. First, co-editor Kaufman asked me to serve as one of the five or so referees who read each of these manuscripts in their initial form. Second, in a somewhat distant past, prior to his ascending to tenure, full professorship, and directorship of a research center, I chaired Larry Daniel's dissertation committee at the University of New Orleans (boy, does reciting these facts make me feel old!).

These Articles and My Views in Context

It might be helpful to readers to frame these three articles, and my own views, within the context of views presented within the literature. Certainly at one extreme some authors (cf. Carver, 1978; Schmidt, 1996) have argued that statistical significance tests should be banned from publications. For example, Rozeboom (1997) recently argued that:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students . . . [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism . . . (p. 335)

Schmidt and Hunter (1997), virulent critics of statistical significance testing, similarly argued that, "Statistical significance testing retards the growth of scientific knowledge; it *never* makes a positive contribution" (p. 37, emphasis added).

At the other extreme (cf. Cortina & Dunlap, 1997; Frick, 1996), Abelson (1997) argued that, "Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented" (p. 118). Similarly, Harris (1997) argued that

> Null hypothesis significance testing (NHST) as applied by most researchers and journal editors can provide a very useful form of social control over researchers' understandable tendency to "read too much" into their data . . . [E]ven NHST alone would be an improvement over the current lack of attention to sampling error. (pp. 145, 164)

Some of these defenses of statistical tests have been thoughtful, but others have been flawed (Thompson, 1998b).

Bruce Thompson is a professor and distinguished research scholar in the Department of Educational Psychology at Texas A & M University. He is also an adjunct professor of community medicine at the Baylor College of Medicine. Correspondence regarding this article should be addressed to Bruce Thompson, Department of Educational Psychology, Texas A & M University, College Station, TX 77843-4225 or by e-mail to e100bt@tamvm1.tamu.edu. Related reprints and the author can be accessed on the Internet via URL: **"http://acs.tamu.edu/~bbt6147/"**.

I see Nix and Barnette (1998) as somewhat approaching the Carver (1978)/Rozeboom (1997) end of the continuum. They "believe that most statisticians would [and seemingly should] welcome orderly change that would lead to abandonment of NHST." The authors feel constrained from supporting a ban, not on the merits, but only because of concerns regarding "democratic principles" and "censorship and infringement on individual freedoms."

McLean and Ernest (1998) believe that "our recommendations reflect a moderate mainstream approach." Certainly, their views are intellectually "moderate." A call that their views are "mainstream" requires a factual judgment as regards a moving target – our moving discipline. McLean and Ernest (1998) suggest that tests of statistical significance "must be accompanied by judgments of the event's practical significance and replicability."

I also see Daniel's (1998) views as being moderate, though they may tend a bit more toward the Carver (1978)/Rozeboom (1997) end of the continuum. Thus, the three articles do not include advocacy that the status quo is peachy-keen, and that no changes are warranted (a deficiency that will doubtless be corrected via additional commentaries).

My own views are fairly similar to those of McLean and Ernest (1998) and Daniel (1998). That is, on numerous occasions I certainly have pointed out the myriad problems with rampant misuse and misinterpretation of statistical tests.

However, I have never argued that statistical significance tests should be banned. If I felt these tests were intrinsically evil, as an editor of three journals, I necessarily would have written author guidelines proscribing these tests. And as an author I would also never report p values.

Instead, I generally find statistical tests to be largely irrelevant. Like Cohen (1994), I do not believe that p values evaluate the probability of what we want to know (i.e., the population). Rather, we assume the null hypothesis describes the population, and then evaluate the probability of the sample results (Thompson, 1996).

I am especially disinterested in statistical tests when what Cohen (1994) termed "nil" null hypotheses are used, particularly when testing reliability or validity coefficients. Daniel (1998) makes some excellent points here. We expect reliability and validity coefficients to be .7 or .8. As his table shows, with a n of 10 or 15, we will always attain statistical significance even for minimally acceptable reliability and validity coefficients, so what is the value of such tests with these or larger sample sizes? Abelson (1997) put the point fairly clearly:

> And when a reliability coefficient is declared to be nonzero, that is the ultimate in stupefyingly vacuous information. What we really want to know is whether an estimated reliability is .50'ish or .80'ish. (p. 121)

Thus, editorial policies of *Educational and Psychological Measurement* proscribe use of statistical testing of reliability and validity coefficients, if (and only if) "nil" nulls are used for this purpose.

I believe that evidence of result replicability is very important and is ignored by those many people who do not understand what statistical tests do (e.g., believe that their tests evaluate the probability of the population). Daniel (1998) at one point says, "Statistical significance simply indicates the probability that the null hypothesis is true in the population" (a view I do not accept), but says later that these tests answer the question, "*If* the sample represents the population, how likely is the obtained [sample] result?" (a view I do endorse).

Empirical studies consistently show that many researchers do not fully understand the logic of statistical tests (cf. Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Similarly, many textbooks teach misconceptions regarding these tests (Carver, 1978; Cohen, 1994).

More than anything else, I especially want to see authors always report effect sizes. I concur with the views of McLean and Ernest (1998), who noted that, "In reviewing the literature, the authors were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report." Kirk (1996) and Snyder and Lawson (1993) present helpful reviews of the many types of effect sizes that can be computed.

Regarding effect sizes, some (cf. Robinson & Levin, 1997) have argued that we must always first test statistical significance, and if results are statistically significant, "only if so: (2) effect size information should be provided" (Levin & Robinson, in press).

In Thompson (in press-b) I used a hypothetical to portray the consequences of this view. Two new proteins that suppress cancer metastasis and primary tumor growth in mice are discovered. Two hundred teams of researchers begin clinical trials with humans. Unfortunately, the 200 studies are underpowered, because the researchers slightly overestimate expected effects, or perhaps because the researchers err too far in their fears of "over-powering" (Levin, 1997) their studies. Low and behold, all 200 studies yield noteworthy "moderate" effects for which $p_{\text{CALCULATED}}$ values are all .06.

> [A]m I to understand that these moderate effect sizes involving a pretty important criterion variable may not permissibly be discussed or even reported? . . . In the Thompson world, . . . [i]n this happy example, considerable direct replication evidence is available, so the noteworthy effect is interpreted even though none (zero, nada) of the 200 results is statistically significant. Thus, this is a world in which, in at least some cases, 'surely, God loves the .06 nearly as much as the .05' level of statistical significance (Rosnow & Rosenthal, 1989, p. 1277). (Thompson, in press-b)

Effect Size Reporting

Nix and Barnette (1998) cite others in suggesting that "studies today are more likely to report effect sizes," perhaps because the APA (1994) publication manual "encourages" (p. 18) such reports. However, McLean and Ernest (1998, emphasis in original) diametrically disagree, arguing that "encouraging" effect size reporting "has *not* appreciably affected actual reporting practices," and then cite five *empirical* studies corroborating their views.

Most regrettably, I believe that the pessimistic views of McLean and Ernest (1998) are correct. Indeed, let me cite five additional *empirical* studies of journal reporting practices that present similar findings (Keselman et al., in press; Lance & Vacha-Haase, 1998; Ness & Vacha-Haase, 1998; Nilsson & Vacha-Haase, 1998; Reetz & Vacha-Haase, 1998). In fact, Keselman et al. (in press) concluded that, "as anticipated, effect sizes were almost never reported along with p-values."

I have offered various reasons why the APA "encouragement" has been such a failure. First, an "encouragement" is too vague to enforce. Second, the APA policy

> presents a self-canceling mixed-message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't." (Thompson, in press-b)

Of course, mindless adherence to old habits may also partly explain the glacial movement of the field, because "changing the beliefs and practices of a lifetime . . . naturally . . . provokes resistance" (Schmidt & Hunter, 1997, p. 49). As Rozeboom (1960) observed nearly 40 years ago, "the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

It is my view (Thompson, 1998a; Vacha-Haase & Thompson, 1998) that most authors will simply not change their practices until editorial policies *require* them to do so. These three sets of authors cite three editorial policies (Heldref Foundation, 1997; Murphy, 1997; Thompson, 1994) requiring effect size reporting. Here are some additional editorial policies on this point. [Should *RESEARCH IN THE SCHOOLS* adopt such a policy? Hint. Hint.]

The editor of the *Journal of Consulting and Clinical Psychology* noted in passing that effect sizes are required in that journal, and furthermore that

> Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the *required* effect size but also a consideration of clinical significance. That is, . . . it is also important for the evaluator to consider the degree to which the outcomes are clinically significant (e.g., normative comparisons). . . . A treatment that produces a significant reduction in depressed mood must also be examined to determine whether the reduction moved participants from within to outside the defining boundary of scores for depression. (Kendall, 1997, p. 3, emphasis added)

The editor of the *Journal of Educational Psychology* called for "the provision of both strength-of-relationship measures and 'sufficient statistics' (the latter to permit independent confirmation of a study's statistical findings; statistical power calculations, and access to relevant information for meta-analyses, among others)" (Levin, 1995, p. 3).

The editor of the *Journal of Family Psychology* argued that, "In addition, reporting clinical significance . . . as opposed to mere statistical significance would also make treatment research more relevant to practitioners" (Levant, 1992, p. 6). Finally, the editor of the *Journal of Experimental Psychology: Learning, Memory, and Cognition* argued that

> In reporting results, authors should still provide measures of variability and address the issue of the generalizability and reliability of their empirical findings across people and materials. There are a number of acceptable ways to do this, including reporting MSEs and confidence intervals and, in case of within-subject or within-

> items designs, the number of people or items that show the effect in the reported direction. (Neeley, 1995, p. 261)

Highlights of the Three Articles

The three articles each had highlights that particularly appealed to me. For example, Nix and Barnette (1998) present a nice albeit short review of the controversies between Fisher as against Neyman and Pearson, which were never effectively resolved (the consequence of this failed resolution being the hodge-podge of practices we see today). I very much liked their statement, "The *p* value tells us nothing about the magnitude of significance nor does it tell us anything about the probability of replication of a study." As I have noted elsewhere,

> The calculated *p* values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because *p* values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{\text{CALCULATED}}$, and 100 studies with the same single effect size could each have 100 different values for $p_{\text{CALCULATED}}$. (Thompson, in press-a)

Daniel (1998) does a nice job of presenting older quotations to illustrate that we have been haunted by these controversies virtually since the inception of statistical tests. I particularly liked his citation of Berkson, arguing in 1938 that testing significance when the *n* is 200,000 is not very enlightening!

Daniel's (1998) review of editorial policies and how they are applied was also informative. He emphasizes a point that some authors do not appreciate: editors will not accept articles that violate their published editorial policies, so prudent authors must take these policies seriously. I find myself in general agreement with Daniel's (1998) very specific recommendations for improving our scholarship.

As regards McLean and Ernest (1998), I very much appreciated their recognition that science is subjective and that statistical tests cannot make it otherwise (Thompson, in press-c). I also very much liked their treatment of the "language controversy."

McLean and Ernest (1998) prefer to keep statistical tests within the researcher's arsenal but are more than willing to provide both effect size and replicability evidence of one or more sorts. I am somewhat less interested than they in the results of statistical tests, but science will move forward to the extent that the latter two issues are finally seriously considered within our inquiry.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods, 2*, 161-172.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2) 23-32.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Mahwah, NJ: Erlbaum.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology, 15*, 3-5.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Lance, T., & Vacha-Haase, T. (1998, August). *The Counseling Psychologist: Trends and usages of statistical significance testing*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Levant, R. F. (1992). Editorial. *Journal of Family Psychology, 6,* 3-9.

Levin, J. R. (1995). Editorial: Journal alert! *Journal of Educational Psychology, 87,* 3-4.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12,* 84-106.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement, 58,* 311-331.

Levin, J. R., & Robinson, D. H. (in press). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review.*

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2)15-22.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82,* 3-5.

Neeley, J. H. (1995). Editorial. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 261.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299-1301.

Ness, C., & Vacha-Haase, T. (1998, August). *Statistical significance reporting: Current trends and usages within Professional Psychology: Research and Practice.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Nilsson, J., & Vacha-Haase, T. (1998, August). *A review of statistical significance reporting in the Journal of Counseling Psychology.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2)3-14.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Reetz, D., & Vacha-Haase, T. (1998, August). *Trends and usages of statistical significance testing in adult development and aging research: A review of Psychology and Aging.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rosenthal, R. & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology, 55,* 33-38.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, *57,* 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Snyder, P. A., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61,* 334-349.

Thompson, B. (Guest Ed.). (1993). Special issue on statistical significance testing, with comments from various journal editors. *Journal of Experimental Education, 61*(4).

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998a, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas.* Invited address presented at the annual meeting of the American Educational Research Association, San Diego.

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799-800.

Thompson, B. (1998c). Review of *What if there were no significance tests?* by L. Harlow, S. Mulaik & J. Steiger (Eds.). *Educational and Psychological Measurement, 58,* 332-344.

Thompson, B. (in press-a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*. [Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago.]

Thompson, B. (in press-b). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Thompson, B. (in press-c). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology*.

Vacha-Haase, T., & Thompson, B. (1998, August). *APA editorial policies regarding statistical significance and effect size: Glacial fields move inexorably (but glacially)*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 39-41

Comments on the Statistical Significance Testing Articles

Thomas R. Knapp
The Ohio State University

This review assumes a middle-of-the-road position regarding the controversy. The author expresses that significance tests have their place, but generally prefers confidence intervals. His remarks concentrate on ten errors of commission or omission that, in his opinion, weaken the arguments. These possible errors include using the jackknife and bootstrap procedures for replicability purposes, omitting key references, misrepresenting the null hypothesis, omitting the weaknesses of confidence intervals, ignoring the difference between a hypothesized effect size and an obtained effect size, erroneously assuming a linear relationship between p and F, claiming Cohen chose power level arbitrarily, referring to the "reliability of a study," inferring that inferential statistics are primarily for experiments, and recommending "what if" analyses.

Since I take a middle-of-the-road position regarding the significance testing controversy (I think that significance tests have their place, I generally prefer confidence intervals, and I don't like meta-analysis!), I would like to concentrate my remarks on ten errors of commission or omission that in my opinion weaken the arguments in these otherwise thoughtful papers. In this article, the three articles under review are referred to as Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998).

1. Each of the authors discusses something they call "internal replicability analysis." The term is apparently due to Thompson (1994), and it represents a misinterpretation of the work on the jackknife and the bootstrap in the statistical literature. I challenge all of the authors to find in that literature (e.g., Diaconis & Efron, 1983; Efron & Gong, 1983; Mooney & Duval, 1993; Mosteller & Tukey, 1977) any reference to either approach providing evidence for the replicability of a finding. They are simply procedures for estimating sampling error without making the traditional parametric assumptions. The confusion may arise from the fact that both require the creation of several replications of the statistic of principal interest (the jackknife by "re-sampling" the sample data without replacement; the bootstrap by "re-sampling" the data with replacement).

2. None of the authors cite the article by Abelson (1997), and two of the authors (McLean and Ernest (1998) and Nix and Barnette (1998)) do not even cite the book on the significance testing controversy (Harlow, Mulaik, & Steiger, 1997) in which that article appears. It is the best defense of the use of significance tests I have ever read. Since the controversy has been going on for many years it is impossible to cite every relevant source, but McLean and Ernest (1998) don't even cite Schmidt (1996), the most vocal critic of significance tests and strongest advocate of meta-analysis. Daniel (1998) cites Thompson's (1998) review of the Harlow et al. compendium, but does not cite Levin's (1998) review that appeared in the same source.

3. Two of the authors make mistakes when discussing what a null hypothesis is. Daniel (1998) gives an example where the null hypothesis is said to be: r (the sample r) is equal to zero, and claims that "by definition" a test of significance tests the probability that a null hypothesis is true (the latter is OK in Bayesian analysis but not in classical inference). Both Daniel (1998) and Nix and Barnette (1998) refer to the null hypothesis as the hypothesis of no relationship or no difference; no, it is the hypothesis that is tested, and it need not have zero in it anyplace.

4. None of the authors point out the weaknesses of confidence intervals or how they can be misinterpreted just as seriously as significance tests. For example, it is not uncommon to see statements such as "the probability is .95 that the population correlation is between a and b." A population correlation doesn't have a probability and it is not "between" anything; it is a fixed, usually unknown, parameter that may be bracketed or covered by a particular confidence interval, but it doesn't vary.

5. None of the authors make sufficiently explicit the necessary distinction between a hypothesized effect size and an obtained effect size. It is the former that is relevant in determining an appropriate sample size; it is the latter that provides an indication of the "practical significance"

Thomas R. Knapp is a professor of nursing and education at The Ohio State University. Correspondence regarding this article should be addressed to Thomas R. Knapp, College of Nursing, The Ohio State University, 1585 Neil Avenue, Columbus, OH 43210-1289 or send e-mail to knapp.5@osu.edu.

of a sample result and around which a confidence interval can be constructed. Cohen (1988) at least tried to differentiate the two when he put the subscript s on the d for the obtained effect size. Some of the confusion in the significance testing controversy could be avoided if we had different terms for those two kinds of "effect sizes." (A similar confusion has arisen recently regarding prospective and retrospective power – see Zumbo & Hubley, 1998.)

6. Daniel (1998) claims that a df of 300 for an ANOVA error term is five times more likely to produce a statistically significant difference than a df of 60. That's not true; the relationship between p and F is not linear.

7. McLean and Ernest (1998) claim that Cohen (1988) recommended a power of .80 as arbitrarily as Fisher recommended an alpha of .05. That's not fair. He (Cohen) argued there, and elsewhere, that Type I errors are generally more serious than Type II errors and therefore beta (= 1 - power) can be chosen to be considerably larger than alpha.

8. Nix and Barnette (1998) refer to "the reliability of the study." There is no such thing as the reliability of a study. Measuring instruments have varying degrees of reliability (I think the claim by Daniel (1998), and others, that reliability pertains to scores, not instruments, is much ado about nothing); statistics have varying degrees of reliability, in the sense of sampling error; studies do not.

9. Nix and Barnette (1998) also seem to suggest that inferential statistics in general and significance testing in particular are primarily relevant for experiments (given their several references to "treatments"). Statistical inference actually gets very complicated for experiments, since it is not clear what the population(s) of interest is (are). Experiments are almost never carried out on random samples, but all true experiments have random assignment. What inference is being made (from what to what) is a matter of no small confusion. (See the reaction by Levin, 1993 to Shaver, 1993 regarding this issue.)

10. Daniel (1998) advocates, as does Thompson, "what if" analyses (not to be confused with the "What if . . . ?" title of the Harlow book). Although such analyses are not wrong, they are unlikely to be very useful. Researchers have actual sample sizes and actual values for their statistics; speculating as to what might have happened if they had bigger or smaller sample sizes, or the population correlations had been bigger or smaller, or whatever, is the sort of thinking that should be gone through before a study is carried out, not after. (See Darlington, 1990, pp. 379-380 regarding this matter.)

But to end on a positive note, I commend Daniel (1998) for his point that a significance test tells you nothing about the representativeness of a sample; McLean and Ernest (1998) for their contention that significance tests (*and* confidence intervals?) aren't very important for huge sample sizes; and Nix and Barnette (1998) for bringing to the attention of the readers of this journal that there are both significance tests and confidence intervals available for multivariate analyses. Curiously, most of the controversy about significance testing has been confined to univariate and bivariate contexts.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would have to be invented). In L.L Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd. ed.). Hillsdale, NJ: Erlbaum.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS*, *5*(2), 23-32.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*, 116-130.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36-48.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Levin, J. R. (1998). To test or not to test H_0. *Educational and Psychological Measurement*, *58*, 311-331.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *RESEARCH IN THE SCHOOLS*, *5*(2), 15-22.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage.

Mosteller, F., &: Tukey, J.W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A Review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS*, *5*(2), 3-14.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Shaver, J. P. (1993) What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*, 293-316.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, *62*, 157-176.

Thompson, B. (1998). Review of *What if there were no significance tests?* by L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.). *Educational and Psychological Measurement*, *58*, 332-344.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, *47*, 385-388.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 43-53

What If There Were No More Bickering About Statistical Significance Tests?

Joel R. Levin
University of Wisconsin – Madison

Questions and concerns are directed to those who advocate replacing statistical hypothesis testing with alternative data-analysis strategies. It is further suggested that: (1) commonly recommended hypothesis-testing alternatives are anything but perfect, especially when allowed to stand alone without an accompanying inferential filtering device; (2) various hypothesis-testing modifications can be implemented to make the hypothesis-testing process and its associated conclusions more credible; and (3) hypothesis testing, when implemented intelligently, adds importantly to the story-telling function of a published empirical research investigation.

From the local pubs to our professional "pubs," everyone in social-science academic circles seems to be talking about it these days. Not that there's anything wrong with talking about it, mind you, even to a more practically oriented crowd such as the readership of this journal. But as with the "gates" of Washington politics on the one coast and the Gates of Washington state on the other, when do we stand up and say "Enough already!"? When do we decide that ample arguments have been uttered and sufficient ink spilled for us to stop talking about it and instead start doing something about it?

The "it," of course, is the "significance test controversy" (Morrison & Henkel, 1970), which, in its most extreme form is whether or not conductors/reporters of scholarly research should continue (or even be *allowed* to continue) the time-honored tradition of testing statistical hypotheses. As has been carefully documented in our current forum on the issue in this issue of *RESEARCH IN THE SCHOOLS*, the topic isn't one that just recently arrived on the science scene. Not at all. Eminent statisticians, applied researchers, and just plain folks have been debating the virtues and vices of statistical significance testing for decades, with the debate crescendoing every couple of decades or so – consistent with principles of GC ("generational correctness").

The decade of the 1990s has been a critical one in hypothesis testing's protracted struggle for survival. During this decade especially vitriolic attacks, by especially viable attackers, in especially visible outlets (e.g., Cohen, 1990, 1994; Kirk, 1996; Schmidt, 1996), have been mounted for the greater good of God, country, and no more significance testing! Even more critically for the life-and-death struggle, in the 1990's we also witnessed the first formal establishment of task forces and committees representing professional organizations [e.g., the American Psychological Association (APA), the American Educational Research Association (AERA), the American Psychological Society (APS)] to study the "problem" and make recommendations. As the deliberations of such task forces have proceeded apace, so have the spoken and written words: for example, in semi-civilized debates at professional meetings [e.g., "Significance tests: Should they be banned from APA journals?" (APA, 1996); "Should significance tests be banned?" (APS, 1996); "A no-holds-barred, tag-team debate over the statistical significance testing controversy" (AERA, 1998)] and in the most comprehensive, most indispensable, source on the topic, the edited volume *What if there were no significance tests*? (Harlow, Mulaik, & Steiger, 1997; reviewed by Levin, 1998, and Thompson, 1998).[1]

In the typical argument scenario, hypothesis testing is cast as the "bad guy," the impeder of all scientific progress. The prosecution prosecutes the accused, and then the defense defends. That is the basic approach taken in Harlow et al.'s (1997) four focal chapters ("The Debate: Against and For Significance Testing"), as well as in the recent professional meeting set-to's. As each piece of hypothesis-testing evidence is trotted out for public display, the typical juror-consumer goes through a "good point, that sounds reasonable, I hadn't thought of that" self-dialogue before deciding whether to convict or acquit, or just to quit and retreat to his/her original position on the subject.

Joel R. Levin is a professor of educational psychology at the University of Wisconsin. Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, 1025 W. Johnson St., University of Wisconsin, Madison, WI 53706 (E-mail address: LEVIN@MACC.WISC.EDU).

Comments and Questions Related to the Present Articles

A similar structure and sequence of events are witnessed in the present collection of three essays. The "bad guy, good guy" script is closely followed, with each essay providing informative backgrounding, coherent evidence, and a convincing closing argument in the form of practical suggestions and proposed solutions. At the same time, even though the Editors of *RESEARCH IN THE SCHOOLS* have striven to be impartial and maintain a balance of perspectives here, the fact that two of the essays are clearly hypothesis-testing indictments whereas only one supports the process indicates that the present scales of justice were tipped a priori toward conviction. Given this unfair state of affairs and not knowing in advance the substance of the other critics' critiques, I can be "up front" in my admission of evening out the imbalance with the comments I am about to make.[2]

All the authors of the present articles cite relevant literature in a scholarly fashion and then proceed to make their case. As a reminder of what those cases are: (a) Nix and Barnette (1998) nix hypothesis testing in favor of a number of more thought-to-be informative alternatives to it (the provision of effect sizes, confidence intervals, replication, meta-analyses); (b) Daniel (1998) basically concurs and then goes on to recommend specific journal editorial-policy measures that could be implemented to effect those changes; and (c) McLean and Ernest (1998) disagree with the fundamental assertion about hypothesis testing's inutility, arguing essentially that it has an important "time and place" (Mulaik, Raju, & Harshman, 1997) in the scientist's analytic arsenal.

Although I have found it unwise to argue with people on matters of politics, religion, and their convictions about hypothesis testing, I will nonetheless attempt to do so by commenting on selected specifics in the three focal articles, in no particular order. Included in my comments are a number of questions that the articles evoked, the responses to which I look forward to reading in the authors' rejoinders. With the exception of Nix and Barnette's discussion of "research registries" (which I found to be a useful notion that should be given serious consideration by social scientists), the case against hypothesis testing introduces all the usual suspects. In that the present authors have examined these suspects in a generally commendable fashion, I will do my best to cross-examine them. In addition to being invited to serve as a commentator on these articles, I was encouraged to get in my own "two bits worth." And so I shall, beginning with a confession: Because of my previously professed "pro" position in the hypothesis-testing debates, I apologize in advance for disproportionately carping and sniping more at the "con" positions of Nix-Barnette and Daniel.

Hypothesis-Testing Fever/Furor

Considerable issue can be taken with something that Nix and Barnette claim early on, namely, that "the informed stakeholders in the social sciences seem to be abandoning NHST . . ." As one who considers himself to be an informed stockholder, I'd be curious to learn to whom Nix and Barnette are referring, on what survey or other supporting reference their claim is made, and exactly how prevalent this abandonment is. One has to wonder: If the perniciousness of hypothesis testing is so pervasive, then why has APA's elite task force recommended that the practice *not* be abandoned, but rather supplemented and improved by many of the same enhancements that are mentioned in the present exchange (*viz*., effect magnitude measures, confidence intervals, replications, and meta-analysis)?

It is understandable that much, if not most, of what Daniel decries and prescribes has been decried and prescribed before. It is understandable because: (a) Daniel draws heavily from the words and work of Bruce Thompson (11 references and counting); and (b) Daniel, as a Thompson collaborator (Thompson & Daniel, 1996a, 1996b), is undoubtedly quite familiar with that corpus. Prominent in Daniel's list of hypothesis-testing do's and don'ts are Thompson's (e.g., 1996) "big three" recommended editorial policy "requirements" for authors of empirical studies – namely, that authors must always: (a) modify the word "significant" with "statistically," in reference to hypothesis tests; (b) include explicit effect-size information; and (c) provide some form of outcome "replicability" evidence.

"Significance" Testiness

Such proposed editorial policy changes are sensible enough and I clearly support the spirit – though not the letter – of them (e.g., Levin & Robinson, in press; Robinson & Levin, 1997). What is difficult to support are requirements that take away certain freedoms of author style and expression; in particular, when editorial policy is only half a vowel away from turning into editorial police. For example, when addressing a professional audience with a shared understanding of technical terminology, why should an author be forced into using stilted, reader-unfriendly, language (e.g., "The two correlations are each statistically significant but not statistically significantly different from one another.")? In a Results section where statistical hypotheses are being tested, there can be no misunderstanding what the word "significant" does or does not mean; the context disambiguates the concept. On the other hand, if an author who detects an effect that is significant statistically (e.g., a significance probability of $p = .01$) but insignificant practically (e.g., a standardized difference in means represented by a Cohen's d of .01) goes on to talk about the effect with reckless hyperbole, then, yes, that author should be shot at sunrise – or at least appropriately chastised.[3]

Effect-Size Defects?

Speaking of talking, the just-mentioned confusion represents a profound mismatch between an author's evidence and his/her words, stemming from a preoccupation with statistical significance at the expense of taking into account the magnitude of the obtained effect (which in the $d = .01$ case was minuscule). However, I have problems with the other side of the "nouveau" editorial-policy-recommendations coin regarding effect-size reporting as well. I will mention a few such problems, none of which is noted either by Daniel or by Nix and Barnette.

First, and even though I am all for including effect sizes as ancillary evidence of outcome importance, it has been pointed out previously (Levin & Robinson, in press; Robinson & Levin, 1997) that there are extremists in the mandatory effect-size camp (including journal reviewers and editors) who advocate reporting and concentrating on effect sizes *only* (i.e., without accompanying statistical/probabilistic support). This practice is absurdly pseudoscientific and opens the door to encouraging researchers to make something of an outcome that may be nothing more than a "fluke," a chance occurrence. Without an operationally replicable screening device such as statistical hypothesis testing, there is no way of separating the wheat (statistically "real" relationships or effects) from the chaff (statistically "chance" ones), where "real" and "chance" are anchored in reference to either conventional or researcher-established risks or "confidence levels." McLean and Ernest's description of Suen's (1992) "overbearing guest" analogy is especially apt in this context.[4]

Examples of the seductive power of large observed effect sizes that are more than likely the result of chance outcomes are provided by Levin (1993) and Robinson and Levin (1997). In its extreme form, effect-size-only reporting degenerates to strong conclusions about differential treatment efficacy that are based on comparing a single score of one participant in one treatment condition with that of another participant in a different condition. Or, even more conveniently and economically (i.e., in situations where time and money are limited), how about conclusions from a "what if" meta-experiment in which scores of two *imaginary* participants are compared ($N = 0$ studies)? The latter tongue-in-cheek situation aside, consider the following proposition:

> Suppose that Aladdin's genie (Robin Williams?!) pops out of the lamp to grant you only *one* forced-choice wish in relation to summarized reports of empirical research that you will read for the rest of your lifetime: You can have access to either a statistical-significance indicator of the reported findings or a practical-significance index of them, but not both (and no sample-size information can be divulged). Which would you choose?

Personally speaking, it would be painful to have to choose only one of these mutually exclusive alternatives. Based on the aforementioned "chance" and "seductive effect size" arguments, however, I think that a strong case can be made for statistical over practical significance. McLean and Ernest's chance-importance-replicability trichotomy represents a nice way of thinking about the problem, with an assessment of the findings' nonchanceness and replicability each given priority over importance. At the same time, I heartily endorse Nix and Barnette's statement, "We would like to see a situation where all studies that were adequately designed, controlled and measured would be reported, regardless of statistical significance." In fact, I am quite sympathetic with others who have called for manuscript reviews and editorial decisions based on just a study's rationale, literature review, and methods and procedures, in the form of a research proposal – with the associated outcomes and data analyses not included until an editorial decision has been reached (e.g., Kupfersmid, 1988; Levin, 1997; Walster & Cleary, 1970a).

So you want to change the world? Nix and Barnette, as well as Daniel, make it sound as though the research world will be a far better place when the hypothesis-testing devil is ousted by the effect-size angel. In my opinion, that is not a fair assumption, as effect-size calculating and reporting are subject to the same "bias" criticisms inherent in familiar "how to lie with statistics" treatises. How to lie with effect sizes? Levin and Robinson (in press) have noted how researchers can select from any number of conventional effect-size measures (including both more and less conservative variants of the indices listed in Nix and Barnette's Table 1, among others) to make the preferred case for their own data. Another problem associated with relying on commonly calculated effect sizes alone is illustrated in the following hypothetical example.

Suppose that an investigator wants to help older adults remember an ordered set of ten important daily tasks that must be performed (insert and turn on a hearing aid, take certain pills, make a telephone call to a caregiver, etc.). In a sample of six elderly adults, three are randomly assigned to each of two experimental conditions. In one condition (A), no special task instruction is given; and in the other (B_1), participants are instructed in the use of self-monitoring strategies. Following training, the participants are observed with respect to their success in performing the ten tasks. As can be seen in the first two columns of Table 1, the average number of tasks the participants correctly remembered to perform was 1.33 and 3.33 for the no-instruction (A) and self-monitoring (B_1) conditions, respectively. For the data provided in Table 1, it can be determined that the "conditions" factor accounts for a hefty 82% of the total variation in task performance (i.e., the squared point-biserial correlation is .82, which for the two-sample case, is equivalent to the sample η^2). Alternatively, the self-monitoring mean is 3-½ within-group standard deviations higher than the no-instruction mean (i.e., Cohen's d is 3.5). From either effect-size perspective (η^2 or d), certainly this represents an impressive treatment effect, doesn't it? Or does it?

Table 1
Hypothetical Data Illustrating Equivalent Standardized Effect Sizes (Condition B Versus Condition A) With Vastly Different Practical Implications

	Condition A	Condition B_1	Condition B_2
	1	3	5
	1	3	8
	2	4	10
M	1.333	3.333	7.667
SD	.577	.577	2.517

Suppose that instead of self-monitoring training, participants were taught how to employ "mnemonic" (systematic memory-enhancing) techniques (B_2) – see, for example, Carney & Levin (1998) – with the results as indicated in the third column of Table 1. The corresponding B_2 mean is 7.67 correctly remembered tasks and a comparison with no-instruction Condition A surprisingly reveals that once again, the conditions factor accounts for 82% of the total variation in task performance (equivalently, d again equals 3.5).[5] Thus, when expressed in standardized/relative terms (either η^2 or d), the effect sizes associated with the two instructional conditions (B_1 and B_2) are exactly the same, and substantial in magnitude. Yet, when expressed in absolute terms and with respect to the task's maximum, there are important differences in the "effects" of B_1 and B_2: Increasing participants' average performance from 1.33 to 3.33 tasks remembered seems much less impressive than does increasing it from 1.33 to 7.67. Helping these adults remember an average of only 3 of their 10 critical tasks might be regarded as a dismal failure, whereas helping them remember an average of almost 8 out of 10 tasks would be a stunning accomplishment. Yet, the conventional effect-size measures are the same in each case.[6]

How, then, not to lie with effect sizes? To borrow from Cuba Gooding, Jr.'s character in the film, *Jerry Maguire*: Show me the data! Show me, the reader, "sufficient" data (American Psychological Association, 1994, p. 16) either in raw (preferably) or in summary form. Then, let me, the reader, decide for myself whether a researcher's particular finding is educationally "significant" or "important," with respect to the standards that I regard as "significant" or "important" (see also Prentice & Miller, 1992).

Lack-of-confidence intervals. Briefly noted here are other suggested alternatives to hypothesis testing that are briefly noted by Daniel, as well as by Nix and Barnette. These include the inclusion of confidence intervals and meta-analyses, both of which are signature recommendations of Schmidt and Hunter (e.g., 1997). As far as the former are concerned, it is well known that one can simply slap a standard error and degree of confidence on an effect size and build a confidence interval *that is equivalent to testing a statistical hypothesis* (but see McGrath, 1998). Schmidt, Hunter, and their disciples, however, eschew that particular application and instead encourage researchers to select two or three or four or five degrees of confidence (e.g., 99%, 95%, 90%, 80%, 70%) and then build/display two or three or four or five corresponding confidence intervals. Well and good, but how is the researcher or reader to interpret these varying-degrees-of-confidence intervals, and what is one to conclude on the basis of them (e.g., when a 95% interval includes a zero treatment difference but a 90% interval does not)? How much confidence can one have in such subjective nonsense?

I never met a meta-analysis . . . Concerning meta-analyses: I have nothing against them. They can be extremely valuable literature-synthesis supplements, in fact. Yet, their purpose is surely quite different than that of an individual investigator reporting the results of an individual empirical study, especially when the number of related studies that have been previously conducted are few or none. Alas, what's a poor (graduate-student or otherwise) single-experiment researcher to do (Thompson, 1996)? Of course, if the logical corollary to the meta-analysis argument is that no single-experiment reports should be published in empirical journals as we currently know them, then count me in! I strongly endorse the recommendation that replications and multiple-experiment "packages" comprise an essential aspect of a researcher's LPU ("least publishable unit") – see, for example, Levin (1991, p. 6).

Robust Conclusions Versus Replicated Outcomes

There's something about "replication" in two of the present articles with which I take issue. That something is a restatement of Thompson's (1993) view that data-analysis strategies such as cross-validation, bootstrapping, and jackknifing "indicate the likelihood of replication" (Nix and Barnette) or "may provide an estimate of replicability" (Daniel). For readers not in the know and who might be misled by such semantic twists, allow me to elaborate briefly. A "replication" defined by corroborating analyses based on alternative slices or samples of the same data – which applications of the just-mentioned statistical procedures attempt to do (see, for example, Efron & Gong, 1983) – is nice for establishing *the robustness of a single study's conclusions* (Thompson's "internal" replication). However, that type of "replication" is neither as impressive nor as imperative for the accumulation of scientific knowledge as is a "replication" defined by *an independently conducted study* (i.e., a study conducted at different sites or times, with different specific participants and operations) *that yields outcomes highly similar to those of the original study* (Thompson's "external" replication) – see, for example, Neuliep (1993) and Stanovich (1998). Even to suggest that researchers should be satisfied with the former, by rationalizing about researchers' diminished physical or fiscal resources (as both Thompson and Nix and Barnette do), is not in the best interest of anyone or anything, and especially not in the best interest of educational research's credibility within the larger scientific community.

What if there were no more bickering about significance tests? Conclusion robustness itself is a matter of no small concern for researchers, for outcome "credibility" (Levin, 1994) and generalizability depend on it. Yet, because of the excessive "heat" (Thompson, in press) being generated by hypothesis-testing bickerers, little time is left for shedding "light" on how to enhance the conclusion robustness of educational and psychological research. In addition to the methodological adequacy of an empirical study (e.g., Levin, 1985; Levin & Levin, 1993; Stanovich, 1998), the credibility of its findings is a function of the study's "statistical conclusion validity" (Cook & Campbell, 1979), which in turn encompasses a consideration of the congruence between the statistical tools applied and their associated distributional assumptions. Reviews of the literature indicate that precious little attention is being paid by researchers and journal referees alike to that congruence: Statistical tests are being mindlessly applied or approved even in situations where fundamental assumptions underlying them are likely grossly violated (e.g., Keselman et al., in press; Wilcox, 1997).[7] Bickering time spent on significance testing is also time away from considering other critical conclusion-robustness issues, including particularly those associated with the pervasive educational research realities of: nonindependent sampling, treatment, and testing units; random (as opposed to fixed) treatment factors; longitudinal and other multivariate designs, among others (e.g., Clark, 1973; Kratochwill & Levin, 1992; Levin, 1992a; Raudenbush & Bryk, 1988; Willett & Sayer, 1994). Accompanied or not by significance testing per se, such statistical issues remain properly "significant."

That concludes my comments on the "big issues" addressed by the three focal articles in this issue of *RESEARCH IN THE SCHOOLS*. Before concluding with a few additional big issues of my own, I will address several misleading and erroneous statements that appear in the present articles. Though not of the magnitude of the issues just discussed, these statements are nonetheless sufficiently distressing that they should not go unmentioned.

Misleading and Erroneous Assertions in the Present Articles

It is bad enough when *consumers* of research reports are uninformed with respect to the methods and meanings of the data analyses reported (as has been claimed, for example, with respect to the hypothesis-testing term "significant"). Even worse is when *researchers/authors* are misinformed with respect to those methods or meanings. But worst of all is when *critics* of data-analytic practices dangerously mislead or make erroneous assertions regarding those practices – and particularly when the words "misuse and misinterpretation" are featured in the title of a critic's critique (as in Daniel's article, for example).

Sample size and statistical power. To wit, consider Daniel's comments about the components of an *F*-test of mean differences, which I quote [with numbers added for convenience in referencing]:

> . . . the mean square for the error term will get smaller as the sample size is increased [1] and will, in turn, serve as a smaller divisor for the mean square for the effect [2], yielding a larger value for the *F* statistic [3]. In the present example (a two-group, one-way ANOVA), a sample of 302 would be five times as likely to yield a statistically significant result as a sample of 62 simply due to a larger number of error degrees of freedom (300 versus 60) [4].

What a misrepresentation of the *F*-test and its operating characteristics! The error mean square (*MSE*) is an unbiased estimator of the population variance (σ^2) that is not systematically affected by sample size. What increasing sample size does is to reduce the sampling variability associated with each condition's mean, which results in increased variability *among* those means, which in turn increases the mean square between conditions (*MSB*) in the *F*-test's *numerator*. Propositions [1] and [2] are therefore false, which invalidates proposition [3]. Proposition [4] is not true as a result of the preceding illogic.

It is also false as a consequence of Daniel's stated "larger number of error degrees of freedom." Again, larger sample sizes increase statistical power by decreasing the sampling variability associated with each condition's mean, which operates to increase the variability among those means. None of this works automatically to increase the *F*-statistic by a constant amount, however, as is asserted by Daniel (e.g., "by five times"), *unless* it is also stated that *all else (except sample size) is held constant* – which includes the value of *MSE* and the means for each condition (all of which are statistics that will vary unsystematically with changes in sample size). To give the impression that merely increasing sample size *guarantees* a larger *F*-ratio, as Daniel and others imply, is unfortunate because it simply is not true.

Show you the data? Don't press the issue. I could come up with dozens – if not hundreds, thousands, or zillions, if I had the time and temperament – of examples from actual empirical studies, many from my own substantive research program, where an *F*-ratio based on small sample sizes (calculated, for example, early in the data-collection process) becomes *smaller* when based on larger or final sample sizes.

Some of Nix and Barnette's assertions about statistical power and a study's publishability are similarly misleading. First, the authors state that the problem is of special concern in educational research, where "... effect sizes may be subtle, but at the same time, may indicate meritorious improvements in instruction and other classroom methods." If instructional improvements are indeed "meritorious," then: (a) effect sizes will not be "subtle;" and (b) even with modest sample sizes, statistical significance will follow. Second, many readers are likely to be misled by the authors' statements that "reliability . . . can be controlled by reducing . . . sampling error" and "the most common way of increasing reliability . . . is to increase sample size." Reducing *sampling* error or increasing *sample* size (the number of participants) does not increase reliability. Reducing *measurement* error or increasing *test* size (the number of items) does. Increasing sample size increases the power or sensitivity of a statistical test, however.

Errors and effect sizes. Nix and Barnette also state that in a hypothesis-testing context, "errors can be due to treatment differences." This statement will come as news to many and deserves some elaboration. In the section entitled "Misunderstanding of *p* values," the authors caution that "differences of even trivial size can be judged to be statistically significant when sampling error is small (due to a large sample size and/or a large effect size) . . ." How can a difference be simultaneously "trivial" and "large?" Read that sentence again. Later in the same section, the authors argue that researchers should "continue to determine if the statistically significant result is due to sampling error or due to effect size." The imprecisely worded statement may lead an uninitiated reader to believe that it is actually possible for a researcher to make such a precise either-or determination, when it is not. In Nix and Barnette's section, "Interpreting effect size," the impression is given that the various U measures are separate/unrelated, when in fact they are alternative ways of thinking about the same outcome – just as is converting *d* (a standardized difference in means) to *r* (the correlation between treatment and outcome), something that was left unsaid. Omitted from a subsequent paragraph is the caution that comparing single-study effect sizes with composite effect sizes can be grossly misleading unless all treatments in question are evaluated relative to functionally equivalent "control" groups (see also Levin, 1994).

Hypothesis Testing as a Meaningful, Memorable Process

In this section I will provide a few personal thoughts about statistical hypothesis testing and its rightful role in the analysis and reporting of empirical research in education and psychology.

Dump the Bathwater, Not the Baby...

No, statistical hypothesis testing, as is generally practiced, is not without sin. I too oppose mindless (e.g., Cohen's, 1994, "rare disease" scenario; Thompson's, 1997, "reliability/validity coefficient testing" criticism) and multiple (e.g., testing the statistical significance of all correlations in a 20 x 20 matrix) manifestations of it. Such manifestations surely portray the practice of hypothesis testing at its worst. More forethought and restraint on the part of researchers would likely help to deflect much of the criticism concerning its misapplication.

Absent in each of the present articles' proposed replacement therapies for traditional statistical hypothesis testing are *alternative hypothesis-testing therapies themselves* – which I have referred to generically as "intelligent" hypothesis-testing practices (Levin, 1995) and which have been articulated in a set of ideal principles

(Levin, 1998). The overarching premise is that statistical hypothesis testing can be a valuable decision-making tool, if implemented in conjunction with a researcher's a priori (i.e., prior to data collection) planning, specification, or determination of:

- a select number of carefully developed (preferably, theory-based) hypotheses or predictions
- a statistical test or tests that validly and parsimoniously assess those hypotheses
- Type I error probabilities that are adequately controlled
- magnitudes of effects that are regarded as substantively "important," along with their associated probabilities of detection
- magnitudes of effects that are regarded as substantively "trivial," along with their associated probabilities of nondetection
- sample sizes that directly follow from these specifications.

The more of these ingredients that are incorporated into the hypothesis-testing process, the more intelligent and informative is that process.

Effects that emerge as statistically significant as a result of intelligent hypothesis testing should be supplemented by ancillary "practical significance" information, including effect sizes (based on relative and/or absolute metrics), confidence intervals, and even – heaven forbid! – more "qualitative" assessments of treatment efficacy (e.g., experimenter observations and participant self-reports). *The* most important supplement to this statistical basis for scientific hypothesis confirmation is evidence accumulation, initially through empirical replications (Levin's, 1995, "A replication is worth a thousandth *p*-value.") and ultimately through literature syntheses (which include the tools of meta-analysis).

In contrast to the anti-hypothesis-testing reforms in the graduate-level statistics courses taught at Michigan State (alluded to by Nix and Barnette), UW-Madison colleague Ron Serlin and I attempt to impart intelligent hypothesis-testing practices to our students. In addition to simply teaching and writing about the potential of such improvements to statistical hypothesis testing (e.g., Levin, 1985, 1997; Seaman & Serlin, in press; Serlin & Lapsley, 1993), we also attempt to practice these preachings in our substantive research investigations. For example, Ghatala and Levin (1976, Exp. 2) adapted Walster and Cleary's (1970b) procedure for determining "optimal" sample sizes to distinguish between substantively important and trivial effects based on acceptable Type I error control and statistical power. Similarly, I convinced a former student to incorporate components of "predicted pattern testing" (Levin & Neumann, in press) to provide stronger, more sensible, tests of his theoretically based predictions – see Neumann and DeSchepper (1991, Exp. 3).

To present a case for a place for intelligent statistical hypothesis testing in educational research, I invite you to imagine the following seemingly far-from-educational-research situation:

> Suppose that you are a medical doctor, whose life work is to keep people alive. A particular patient fits a profile for being "at risk" for developing some dangerous abnormality. You need to make a decision, based on a simple screening test, whether or not to proceed to more extensive/expensive testing. For patients with this kind of "at risk" profile, the screening test is known to have a 90% chance of identifying those who have the abnormality to some substantial degree, a 5% chance of identifying those who have the abnormality only to some very minimal degree, and a 1% chance of identifying those who do not have the abnormality at all.[8]

Based on the preceding information, does it seem reasonable to you, as a responsible doctor, to use the screening test as a basis for making a decision about whether or not to proceed to the next phase of evaluation? It does to me.

> OK, now suppose that you are an educational researcher whose life work is to study ways of improving the academic performance of "at risk" students. You have developed a literature-guided intervention for "at risk" middle-school students and you want to assess its effectiveness by comparing the end-of-year educational achievement of students who receive the intervention and those who do not (randomly determined). If the intervention produces a substantial difference in average achievement between the two groups (operationalized as $d = 1.00$), you want to have a 90% chance of detecting it; if it produces a minimal difference ($d = .25$), you only want a 5% chance of detecting it; and if there is no difference at all ($d = .00$), you are willing to tolerate a risk of 1% of falsely detecting that. Adapting the Walster-Cleary (1990b) approach, for example, indicates that the just-specified parameters and probabilities are satisfied if 32 students are randomly assigned to each of the two conditions (intervention and no intervention).

Based on the preceding information, does it seem reasonable to you, as a responsible educational researcher, to perform a statistical test as a basis for making a decision about the intervention's potential? It does to me – and especially because the situation just described incorporates the earlier listed intelligent hypothesis-testing ingredients. I certainly do not claim this hypothetical educational hypothesis-testing example to represent a detail-by-detail correspondence with the equally hypothetical medical screening-test example. Rather, it constitutes a close enough analogy that takes us through a similarly sensible decision-making process.

. . . And Now the Rest of the Story

I conclude my remarks with a story relevant to our discussion of hypothesis testing's proper place on the empirical research plate.

> It is a dark and stormy night. A shot rings out in the presidential palace. A body slumps and falls to the ground, dead. A one-armed man is seen fleeing the scene. Inspectors Poirot and Clouseau are called in to investigate. Poirot determines that the deceased is the president's lover. Clouseau notices a charred sheet of paper in the fireplace. He picks it up. "Ooooohh, it's still hot!" he yelps, but is nonetheless able to discern some scribblings on the paper. "Zoot, alors, I have it! And I know precisely how it happened!" Clouseau crows. He continues: "The murderer is . . . [pause] . . . the president's men . . . [pause] . . . or possibly it's the one-armed man . . . [pause] . . . or perhaps it's even the president herself . . . [pause] . . . I haven't a clew!"

Hey, c'mon, who dunnit? Tell us the rest of the story. Inquiring minds want to know!

So you want to know the ending? Let me tell you a different story. Somewhere along the academic trail I had an epiphany about reports of empirical research in scholarly journals (at least those in the fields of psychology and education): In addition to describing what was done, how it was done, and what was found, a journal article should "tell a story." I'm not using "story" in the fictional sense here, but rather as true to life and justifiable on the basis of the study's specific operations and outcomes. Telling a story, with a clever "hook" and memorable take-home message, represents a key landmark on the publication highway (e.g., Kiewra, 1994; Levin, 1992b; Sternberg, 1996). It is something that editors usually demand, reviewers seek, and readers require. A study without a meaningful, memorable story is generally a study not worth reporting. In certain situations, and in light of my earlier comments, incorporating one or more additional experiments into a one-experiment study often helps to breathe life into an otherwise moribund article.

Exactly what does any of this have to do with our current hypothesis-testing discussion? I believe that an invaluable, though heretofore overlooked, function of statistical hypothesis testing (especially if implemented intelligently) is to assist an author in developing an empirical study's story line and take-home message. Just as with the preceding Clouseauian fantasy with its inconclusive conclusion (or its invent-your-own ending), an empirical research article without an evidence-based conclusion is not likely to satisfy either the reader's affective (interest, enjoyment) or cognitive (understanding, memory) processes. We human animals seek to extract some form of order from the chaos in the world around us; we are all "meaning makers." As consumers of scientific research, we seek to do the same from the jumble of theory, methods, and results that are provided in a journal article. In my opinion, selective, planful statistical hypothesis testing can help one extract order from chaos, not just in the "chance-finding filtering" sense, but in the sense of cementing as firm a conclusion as can be made from the evidence presented until a critical replication-attempting study comes along. I additionally believe that hypothesis testing is much better suited to that cementing task than are other proposed individual alternatives for summarizing the results of single-study investigations, including the provision of effect sizes (are they real?) and multiple-confidence-level confidence intervals (which one do you prefer?).[9]

I could go on about the story-telling function of journal articles and hypothesis testing, but I think you get the idea. As for stories, what's the take-home message of *this* article? There are actually three take-home messages, each enumerated in the Abstract. If you're interested, go back and (re)read them. That, of course, is what journal abstracts are supposed to summarily convey: the "bottom line" of one's work.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

Carney, R. N., & Levin, J. R. (1998). Mnemonic strategies for adult learners. In M. C. Smith & T. Pourchot (Eds.), *Adult learning and development: Perspectives from educational psychology* (pp. 159-175). Mahwah, NJ: Erlbaum.

Clark, H. H. (1973). The language-as-fixed-effects fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.

Daniel, L.G. (1998). Statistical significance testing: A Historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS, 5*(2), 23-32.

Derry, S., Levin, J. R., & Schauble, L. (1995). Stimulating statistical thinking through situated simulations. *Teaching of Psychology, 22*, 51-57.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37*, 36-48.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23*, 89-105.

Frick, R. W. (1995). *Using statistics: Prescription versus practice*. Unpublished manuscript, Department of Psychology, State University of New York at Stony Brook.

Ghatala, E. S., & Levin, J. R. (1976). Phenomenal background frequency and the concreteness/imagery effect in verbal discrimination learning. *Memory & Cognition, 4*, 302-306.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), *Review of Research in Education* (Vol. 5, pp. 351-379). Itasca, IL: Peacock.

Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*.

Kiewra, K. A. (1994). A slice of advice. *Educational Researcher, 23*(3), 31-33.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New developments for psychology and education*. Hillsdale, NJ: Erlbaum.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635-642.

Levin, J. R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C. J. Brainerd (Eds.), *Cognitive learning and memory in children* (pp. 205-233). New York: Springer-Verlag.

Levin, J. R. (1991). Editorial. *Journal of Educational Psychology, 83*, 5-7.

Levin, J. R. (1992a). On research in classrooms. *Mid-Western Educational Researcher, 5*, 2-6, 16.

Levin, J. R. (1992b). Tips for publishing and professional writing. *Mid-Western Educational Researcher, 5*, 12-14.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*, 378-382.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review, 6*, 231-243.

Levin, J. R. (1995, April). *The consultant's manual of researchers' common stat-illogical disorders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12*, 84-106.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement, 58*, 313-333.

Levin, J. R., & Levin, M. E. (1993). Methodological problems in research on academic retention programs for at-risk minority college students. *Journal of College Student Development, 34*, 118-124.

Levin, J. R., & Neumann, E. (in press). Testing for predicted patterns: When interest in the whole is greater than in some of its parts. *Psychological Methods*.

Levin, J. R., & Robinson, D. H. (in press). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53*, 796-797.

McLean, J. E., & Ernest, J. M. (1998). The Role of statistical significance testing in educational research. *Research in the Schools*, *5*(2), 15-22.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). *What if there were no significance tests*? (pp. 65-115). Mahwah, NJ: Erlbaum.

Neuliep, J. W. (Ed.). (1993). Replication research in the social sciences. Special issue of the *Journal of Social Behavior and Personality*, *8*(6).

Neumann, E., & DeSchepper, B. G. (1991). Costs and benefits of target activation and distractor inhibition in selective attention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *17*, 1136-1145.

Nix, T.W., & Barnette, J.J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS*, *5*(2), 3-14.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, *92*, 766-777.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160-164.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, p. 423-475). Washington, DC: American Educational Research Association.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*, 21-26.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. A. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Seaman, M. A., & Serlin, R. C. (in press). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 199-228). Hillsdale, NJ: Erlbaum.

Stanovich, K. E. (1998). *How to think straight about psychology* (5th ed.). New York: Longman.

Sternberg, R. J. (1996). *The psychologist's companion: A guide to scientific writing for students and researchers* (3rd ed.). Cambridge, UK: Cambridge University Press.

Suen, H. K. (1992). Significance testing: Necessary, but insufficient. *Topics in Early Childhood Special Education*, *12*, 66-81.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Education*, 61(4), 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding significance tests: Further comments. *Educational Researcher*, *26*, 29-32.

Thompson, B. (1998). Review of *What if there were no significance tests*? *Educational and Psychological Measurement*, *56*, 334-346.

Thompson, B. (in press). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Thompson, B., & Daniel, L. G. (1996a). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*, 197-208.

Thompson, B., & Daniel, L. G. (1996b). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement*, *56*, 741-745.

Walster, G. W., & Cleary, T. A. (1970a). A proposal for a new editorial policy in the social sciences. *The American Statistician*, *24*, 16-19.

Walster, G. W., & Cleary, T. A. (1970b). Statistical significance as a decision-making rule. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology* (pp. 246-254). San Francisco: Jossey-Bass.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363-381.

Footnotes

[1] The authors of the present exchange can certainly be excused for their limited reference to the Harlow et al.

volume, as it likely was released only after earlier versions of the current articles had been written and submitted.

[2] Psst! It should be a secret to nobody that I am a staunch hypothesis-testing defender (e.g., Levin, 1993, 1998; Robinson & Levin, 1997) – although I do not defend the form in which it is generally practiced. That predilection obviously colors my reactions to the present articles.

[3] As an aside and as not accurately conveyed by McLean and Ernest, we (Levin & Robinson, in press; Robinson & Levin, 1997) do not argue *that* alternative language is needed in Results sections. Rather, we suggest that *if* better language is mandated, then descriptors such as "statistically real," "statistically nonchance," and "statistically different" could readily say what one means and mean what one says without a trace of "significance."

[4] A primary function of statistical hypothesis testing has been analogized in even more colorful terms – a "crap detector" – by a distinguished scholar who shall unfortunately remain nameless in that I cannot locate the appropriate citation at the moment.

[5] In each case, the obtained treatment difference is statistically "real," or nonchance ($p \leq .05$, one-tailed), on the basis of either a parametric or nonparametric hypothesis test.

[6] The major problem in this example arises from the conditions' differing variabilities. That problem could be accounted for by defining alternative *d*-like effect-size measures based on just the control condition's (Condition A's) standard deviation, as has been suggested by Glass (1977), Hedges and Olkin (1985), and others. Interpreting effect sizes, in the absence of raw data, remains a problem for η^2 and Cohen's *d*, however. Concerns about effect sizes based on relative metrics, and a variety of other concerns, are detailed by O'Grady (1982), Frick (1995), and Fern and Monroe (1996).

[7] Note that assumptions violations also affect the validity of other inferential statistical alternatives, such as confidence intervals and meta-analyses. Interestingly and in contrast to the "replication" objectives misattributed to them, bootstrapping and jackknifing are methods that *do* possess either "distribution-free" or other robust qualities that could be exploited to circumvent assumption-violations problems.

[8] In this example, I have tried to mitigate the important "base-rate" problem (e.g., Derry, Levin, & Schauble, 1995) by restricting the population to patients with an "at risk" profile. Even so, the problem remains and would need to be taken into account should the screening test's results prove positive.

[9] On the other hand, if it can be documented that the major impediment to scientific progress lies in the valuelessness of reporting single- or few-study investigations (as some have accused), then why not simply discontinue the production of journals that publish primary-research articles and continue with only those that publish research syntheses? Imagine what a triumph that would be for meta-analysis enthusiasts!

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 55-57

A Review of Hypothesis Testing Revisited: Rejoinder to Thompson, Knapp, and Levin

Thomas W. Nix
University of Alabama

J. Jackson Barnette
University of Iowa

This rejoinder seeks to clarify the authors' position on NHST, to advocate the routine use of effect size, and to encourage reporting results in simple terms. It is concluded that the time for action, such as that advocated in Nix and Barnette's original article, is overdue.

Before we respond to the critiques of our colleagues, we would like to comment that discourse such as that exemplified in this journal issue is the type of debate that is necessary to lead us to more coherent methods of analyzing data. As Mark Twain said, "Loyalty to petrified opinion never yet broke a chain or freed a human soul." The situation we have described (Nix & Barnette,1998) is one that has the potential to mislead those not well versed in statistical methods, the enlightened practitioners who look to educational research for guidance in the most difficult and, in our opinion, the most important of professions, the education of fertile young minds.

Clarification of Our Position

First, we must clarify our position that has been somewhat distorted by the reviews. We do not agree with Schmidt (1996) that Sir Ronald Fisher led us to this point of confusion and chaos in the educational research endeavor (Sroufe, 1997). Fisher deserves praise for bringing to agronomy the methods that have helped agriculture achieve the productivity that we see today. However, Fisher and Pearson allowed their insecurities to seep into their professional lives. Instead of criticizing these great men, we should learn from their human frailties and not allow ourselves to repeat their mistakes.

We do agree with Schmidt that the advancement of knowledge, particularly in educational research, has been stymied by rote adherence to null hypothesis significance testing (NHST). The extensive literature outlining the shortcomings of NHST cannot be ignored; we must look to new methods that will bring more coherence to our field. Our position is not that a draconian ban on NHST should be imposed on the huddled scholarly masses. We agree with Thompson (1998) that NHST's are "largely irrelevant" (p. 5). This is why we have offered alternatives such as effect size measure, confidence intervals, measures of study replicability, meta-analytic studies, and research registries of studies, along with strategies for how we could move in an orderly fashion away from NHST without imposing bans or unnecessary rules.

We do believe that universal standards for social scientific endeavor are in the best interest of advancing knowledge. These standards, after thoughtful study, should apply to scholarly journal submissions, to human use institutional review boards, and to the conduct of meta-analytic studies. Standards, however, should not prohibit the use of any statistical technique. Bans of sacred cows usually only solidify the opposition to rational change. It is our belief that rational change can happen from the top-down through concerted action by the large professional organizations (the APA, AERA, ASA, etc.). What we advocate is not radical change, since models exist in the medical field and in Europe that simulate the actions that we have suggested. The only requirement is action.

Effect Size

Levin (1998) and Knapp (1998) have reported on our enchantment with effect size measures and the methods we advocate. In no way do we mean to imply that these methods are perfect, only better than the existing methods. Cohen (1988) has expressed some of the difficulty in explaining effect size in the multivariate case. Cohen has stated that, ". . . f^2 (the multivariate effect size index) is neither simple nor familiar . . ." (p. 477). Cooper and Hedges (1994) have reported that the early meta-analytic work was "at best an art, at worst a form of yellow journalism" (p.7). All methods have to go through a period of development and expansion. We believe our recommendations would have been foolhardy in the 1970s or 1980s, since the methods we advocate had not gone through rigorous testing. At this point, we believe the period of development is far enough along to advocate the routine use of these methods as a means of advancing social science. In fact, we see further need for empirical research on the relationships among several indicators of treatment influence, including test statistics, *p*-values, confidence intervals, and with effect size measures including eta-squared and omega squared. I (JJB) am particularly interested in how these measures are related and how they are influenced by research design, number of groups and the number of subjects. Yes, effect size and meta-analytic techniques do have their limitations, and we should always remain vigilant to their shortcomings, just as some of our predecessors have with NHST.

Reporting of Results

We do agree with Levin (1998) that writing skill is a necessary prerequisite to good scholarship, but we do not agree that the ability to turn a clever phrase and tell a story should necessarily be part of good writing skill. We would like to see researchers, regardless of their inherent creativity, be able to report valid research results in the simplest terms possible. In this manner not only could researchers understand and appreciate the literature, but practitioners could also glean information from studies that could help in their everyday practice.

A prerequisite to good scholarship and good science is consistency in language. In the world of statistics, this is not a small problem. Vogt (1993) has attempted to explain some of the problems in definitions and vagueness of terms. For example, the symbol β is used to symbolize both the regression coefficient and the probability of a type II error in NHST. Similarly, the intercept and slope in a regression equation are often referred to as constants, when in fact, both have variance and standard errors associated with them. Additionally, researchers often fail to tell readers if the assumptions of a statistical test have been satisfied, let alone even tested, when the lack of adherence to the assumptions confounds the results of many tests. Statisticians understand these problems, but if only statisticians understand research, is the research of any value? For research to be valuable it must be precise and as unambiguous as possible so that is can be comprehended by practitioners as well as other researchers. In this light, as opposed to Levin's preference for statistical significance, we would opt for the practical significance of research over statistical significance.

Apologies and Defense

We must now apologize to Knapp (1998) for our lack of clarity in using the term "reliability" to describe a study (p. 40). We stand corrected on this point. We should have used the term "replicability." However, it should be pointed out that in meta-analytic studies the individual study is a data point. Therefore, in this sense a study could be said to have reliability, if it can be replicated.

Knapp (1998) has also stated that the null hypothesis "need not have zero in it anyplace" (p. 39). In fact the use of H_o: $\mu_1 = \mu_2$ implies that there is no difference in the two population means, or H_o: $\mu_1 - \mu_2 = 0$. As other writers (Bakan, 1966; Cohen, 1988; Hinkle, Wiersma, & Jurs, 1994) have claimed, the null hypothesis is the hypothesis of no difference or no relationship. Of course, it is the hypothesis that is tested, but to say the null hypothesis need not have a zero in it is puzzling.

We agree with Knapp that we used Thompson's (1989, 1993) work as the basis for our recommendation that jackknife and bootstrap methods be used to test (within the limitations of the original data) the replicability of a study without full-scale replication. We also suggested that power of the test could be used as a surrogate for replicability (p. 10) We will leave Thompson (1998) and Knapp (1998) to resolve their disagreement, but conceptually we still believe, no matter what method or indicator is used, that the likelihood of the replicability of a study is important information for the reader and is in the best interest of good science.

Knapp (1998) indicated that we did not reference the outstanding work on the significance testing controversy by Harlow, Mulaik, and Steiger (1997). This is not correct. We reference three chapters that appeared in this book. With regard to Levin's concern about who the "informed stakeholders who are abandoning NHST" are (p. 44), we cited evidence of the first indications of movement away from NHST. Thompson (1998) corrects this assertion by citing sources from 1998 that provide evidence that a shift from NHST to the use of effect size measures is not underway. We stand corrected on this point but must point out that the sources that Thompson cites were unavailable when we developed our arguments.

We appreciate the opportunity to voice our opinions on the state of social science research and the critique of our work. None of our critics have provided sufficient evidence that the advancement of social science would be hampered if authors were required to provide more relevant information in their publications; and we found support for the establishment of research registries to mimic the success that the medical field has had in conducting meta-analyses. Although our ideas are neither unique nor revolutionary, we believe the time for concrete action, such as that we advocate, is long overdue.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6). 423-437.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd ed.). Hillsdale N. J.: Lawrence Erlbaum Associates.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Harlow, L.L., S.A. Mulaik, & Steiger, J.H. (Eds.). (1997). *What if there were no more significance tests?* Mahwah, N.J.; Lawrence Erlbaum Associates.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). *Applied statistics for the behavioral sciences.* (3rd ed.). Boston: Houghton-Mifflin.

Knapp, T. R. (1998). Comments on the statistical significance testing articles. *RESEARCH IN THE SCHOOLS*, *5*(2), 39-41.

Levin, J. (1998). What if there were no more bickering about statistical significance tests? *RESEARCH IN THE SCHOOLS*, *5*(2), 43-53.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS*, *5*(2), 3-14.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115-129.

Sroufe, G.E., (1997) Improving the awful reputation of educational research. *Educational Researcher*, *26*(7), 26-28.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, *22*, 2-6.

Thompson, B (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *6*(4). 361-377.

Thompson, B. (1998). Statistical significance testing and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS*, *5*(2), 33-38.

Vogt, P. W., (1993). *Dictionary of statistics and methodology*. Newbury Park: Sage.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 59-62

Fight the Good Fight: A Response to Thompson, Knapp, and Levin

James M. Ernest
State University of New York at Buffalo

James E. McLean
University of Alabama at Birmingham

After discussing common sentiments in the three papers in this special issue, the authors address concerns of omission expressed by one of the critiquers and provide recommendations for the role of SST.

After reading the three papers (Knapp, 1998b; Levin, 1998; & Thompson, 1998b) that reviewed the articles by Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998), it occurred to us that we "got off" rather lightly. In preparing our response to the contents of the other papers in this special issue of *RESEARCH IN THE SCHOOLS*, we would first like to comment on the general sentiments shared throughout the papers. Secondly, we thought that most of the comments directed toward our paper were concerned with perceived omissions. As Knapp (1998b) pointed out, the controversy has been going on for many years now, and therefore it is impossible to cite every relevant source. However, in this response we will attempt to address Knapp's concerns of omission. Finally, we would like to provide our recommendations for the role of Statistical Significance Testing (SST), agreeing with Thompson (1998b) that the status quo is not "peachy-keen" and that changes are warranted.

Levin (1998) noted that this special issue of *RESEARCH IN THE SCHOOLS* has approached the SST debate as many other forums have regarded the issue. In simple terms (and to use Levin's legal analogy), the debates have cast SST as the "bad guy" of science, often with the hope that the good rational people of the world (or at least those people interested enough to read these journal articles and participate in conferences) may hold trials not so much for, but of, the accused. Unfortunately, the accountability system for SST has not been as favorable as many accountability systems in the world. In the SST accountability system, this accords the accused a status of presumed guilty, and innocence must be proved.

When the topic of SST is raised, it is usually raised in a negative light, the faults of the procedure are considered, and then the issue is opened for proponents of SST to justify the procedure's worth. The debate--before it starts--is stacked against its use. We do not think it would be remiss to say that all people with an interest in the SST debate know there are problems with the SST practice. These problems, according to Hagen (1998), are typically centered around three broad criticisms. The criticisms are concerned with: "(a) the logical foundations of NHST [Null Hypothesis Statistical Testing], (b) the interpretations of NHST, and (c) alternative and supplementary methods of inference" (p. 801). As Hagan (1998) noted, the responses to his 1997 article (Falk, 1998; Malgady, 1998; McGrath, 1998; Thompson, 1998a; & Tryon, 1998) were concerned with all three issues; however, the bulk of Hagan's (1998) response was directed at the logical validity of SST. Rather than our paper being a re-hash of the same arguments concerning the logic of the test, the purpose of our paper was to consider the value of SST as one "of several elements in a comprehensive interpretation of data" (McLean & Ernest, 1998, p. 15).

Our approach to the SST issue was to argue for the positive aspects of SST. We advocated for the use of SST (a limited but necessary use) and also for the necessary inclusion of information concerning the practical significance of the results supported with an index of replicability. As Thompson (1998b) noted, this was a "moderate approach." Also, it was interesting to see Knapp (1998b) refer to his beliefs within a middle-of-the-road position, and Thompson reflect "[m]y own views are fairly similar to those of McLean and Ernest (1998) and Daniel (1998)." When one considers that Levin (1998) confesses to be on the "pro" side in the hypothesis testing debate (with McLean & Ernest, 1998 as pro; Daniel, 1998 and Nix & Barnette, 1998 as con), one realizes that the division between pro and con is not great – one dares to say even "non-significant."

Levin's (1998) reference to the 1998 American Educational Research Association annual meeting session (titled: "A no-holds-barred, tag-team debate over the statistical significance testing controversy") reinforces the idea that there are a number of similarities between those that consider themselves on two sides of a battle. During the debate we saw Tom Knapp and Joel Levin in the "pro" corner, and in the "con" corner were Ron Carver and Bruce Thompson. Yet, even with what seemed to be two diametrically opposed views represented by Carver and Levin, it was interesting to hear Thompson conclude his remarks by stating "I don't think anyone totally disagrees with anyone else."

With respect to Knapp's Comment 1 concerning the challenge to find the idea of "replicability" in the original writings of Mosteller and Tukey (1977), Efron and Gong (1983), Diaconis and Efron (1983), or Mooney and Duval (1993), whom he credits with developing the jackknife and bootstrap procedures: we did not claim that establishing replicability was part of the original purpose of these procedures. We drew the idea from current practice and the writings of Thompson (e.g., 1994). There have been many developments in science and mathematics that have gone far beyond their original purposes. For example, Bonferroni would never have guessed that his inequality would become the basis for numerous multiple comparison procedures. In addition, our recommendation of including an estimate of replicability was not limited to these two approaches. In fact, we believe firmly that the best method of producing support for the replicability of the findings is to replicate the study.

In response to Knapp, the comment that our manuscript omitted the Schmidt (1996) article was well received. However, it is our opinion that the addition of Schmidt's arguments do not add substantially to our original arguments. The main thrust of Schmidt's argument (1996) is to abandon SST and substitute "point estimates of effect sizes and confidence intervals around these point estimates" (p. 116). It should be noted, as Thompson (1998a) advised, that the mindless interpretation of whether the confidence interval subsumes zero is doing nothing more than null hypothesis testing. Thus, Schmidt's rationale for the use of confidence intervals was within the context of comparing multiple studies.

With reference to individual studies, Schmidt's recommendations do not address the possibility of making "something of an outcome that may be nothing more than a 'fluke,' a chance occurrence" (Levin, 1998, p. 45). Another of Schmidt's recommendations is the multiple constructions of confidence intervals, yet as Levin (1998) challenges us, "how is the researcher or reader to interpret these varying-degrees-of-confidence intervals, and what is one to conclude on the basis of them?" (p. 46).

In reflection, with the proliferation of recent articles that address the SST debate, there were many authors' articles omitted. However, within this rejoinder, we felt it appropriate to acknowledge the role of Schmidt within the history of the SST debate. Also, we felt it pertinent to note that we concur with Knapp's (1998a) final summary statement provided during the AERA tag-team debate. Specifically,

> Frank Schmidt, the prime mover in all of this fuss, advocates the discontinuation of ALL significance tests in favor of confidence intervals for single studies and the discontinuation of ALL narrative literature reviews in favor of meta-analyses for synthesizing results across studies. I am pleased to see that he appears to be losing both battles. (Emphasis in original)

Knapp's (1998) comment about Cohen was an interesting point but fails to challenge our initial comment. Knapp noted that we "claim[ed] that Cohen (1988) recommended a power of .80 as arbitrarily as Fisher recommended an alpha of .05." Knapp (1998) continued "[t]hat's not fair. He (Cohen) argued there, and elsewhere, that Type I errors are generally more serious than Type II errors and therefore beta (1 - power) should be chosen to be considerably larger than alpha." We concur, Cohen did argue this point. Also, we agree that Type I errors are generally more serious than Type II errors; however, our issue is that the choice of .80 is just as arbitrary as the choice of .05 for an alpha level. Choosing one number over another (the choice of .05 rather than .06) is an arbitrary matter; choosing .80 rather than .79 is just as arbitrary. These numbers are subjective, and although we agree that the choice of beta should be "considerably larger than alpha" whether one chooses .79 or .80 is arbitrary. With tongue-in-cheek, and in reference to Rosnow and Rosenthal's (1989) comment of "surely God loves the .06 nearly as much as the .05" (p. 1277), surely God loves the .79 nearly as much as the .80 recommendation for power.

In reviewing the research, we feel that a major problem with articles that discuss SST (such as the ones within this special issue of *RESEARCH IN THE SCHOOLS*) is that, more often than not, we are not even "preaching to the choir." It is as though we are preaching to a congregation of ministers. And, more often than not, we are not preaching, we are arguing (or debating what should be a consensus about how we report empirical information). Within our article (McLean & Ernest, 1998) and endorsed by Thompson (1998b), practices have not appreciably affected actual research reporting. When an issue is debated for as long as this issue has been debated, consensus is rare. If an argument is made that statistical testing should be used intelligently (Levin, 1995) including other pertinent pieces of information (an estimate of practical significance, etc.), it would seem reasonable for people to discuss the pros and cons of the issue and come to some consensus.

When statements are made that attack a practice valued by others, such as that NHST "retards the growth of scientific knowledge" (Schmidt & Hunter, 1997, p. 37), nature predicts the initial reaction turning from fright, to flight, to fight. When authors come to the conclusion that "we must abandon the statistical significance test" (Schmidt, 1996, p. 115), or "educational research would be better off without statistical significance testing" (Carver, 1993, 287), researchers who place value in SST fight for the test's validity. Rather than setting up a situation where people "fight the good fight" for their particular beliefs, it would appear prudent to create a situation where it is possible to compromise beliefs. Thus, it is our recommendation that a compromise be made by accepting tests of significance (or not trying to abandon them) and requiring estimates of effect sizes (Thompson, 1998b) along with evidence of external replicability when possible.

Perhaps Suen (1992) said it best: The

> ultimate conclusion of any study and its importance is inherently a human judgement. Significance testing, being mathematical and incapable of making judgements, does not provide such answers. Its role is to filter out the sampling fluctuation hypothesis so that the observed information (difference, correlation) becomes slightly more clear and defined. Judgements can then be more definitive or conclusive. On the other hand, if significance testing fails to filter out the sampling fluctuation hypothesis (i.e., nonsignificance), we may still make our judgement based on the observed information. However, our judgement in this case can never be definitive. (p.79)

As Suen (1992) noted, the value that one may attribute to an empirical study is largely subjective and based on human judgements. Statistics should be viewed as subjective and not, as Abelson (1995) humorously noted, "a set of legal or moral imperatives, such as might be announced at a public swimming pool. (ABSOLUTELY NO DOGS OR FRISBEES ALLOWED. VIOLATORS WILL BE PROSECUTED.)" (p. 56). It is our belief (and in line with Levin's concept of story telling) that the interpretation of statistics should be an exercise of statistical detective work, using as many pieces of the puzzle as possible to inform our decisions.

As noted in our original paper (McLean & Ernest, 1998) and in Levin's response (1998), a case can be made for considering the chance-importance-replicability of empirical findings. This subjective judgement about the utility of the results should be made from as much information as possible. The art of making decisions is exactly that, an art. Ergo, information regarding SST should be included in a research report with at least one measure of practical significance, and if possible (and recommended), evidence of external replication.

Oh, and in reference to Thompson's (1998b) comment that for something to be "mainstream" it requires "a factual judgement as regards a moving target – our moving discipline" (p. 34), Webster's dictionary considers "mainstream" to be a prevailing current or direction of activity or influence. Maybe this was just our wishful thinking.

References

Abelson, R. P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61*(4), 287-292.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS, 5*(2), 23-32.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*, 116-130.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36-48.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53*, 798-799.

Hagan, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.

Hagan, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53*, 801-803.

Knapp, T. R. (1998a, April). *A summary of Tom Knapp's position regarding significance testing*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Knapp, T. R. (1998b). Comments on statistical significance testing articles. *RESEARCH IN THE SCHOOLS, 5*(2), 39-41.

Levin, J. R. (1995, April). *The consultant's manual of researchers' common statillogical disorders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Levin, J. R. (1998). What if there were no more bickering about significance tests? *RESEARCH IN THE SCHOOLS, 5*(2), 43-53.

Malgady, R. G. (1998). In praise of value judgements in null hypothesis testing . . . and of "accepting" the null hypothesis. *American Psychologist, 53,* 797-798.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53,* 796-797.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *RESEARCH IN THE SCHOOLS, 5*(2), 15-22.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods, 1*(2), 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Suen, H. K. (1992). Significance testing: Necessary but insufficient. *Topics in Early Childhood Special Education, 12(1)*, 66-81.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62*, 157-176.

Thompson, B. (1998a). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799-800.

Thompson, B. (1998b). Statistical significance and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS, 5*(2), 33-38.

Tryon, W. W. (1998) The inscrutable null hypothesis. *American Psychologist, 53*, 796.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 63-65

The Statistical Significance Controversy Is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin

Larry G. Daniel
University of North Texas

A rejoinder is offered on the three reviews of Daniel's article (this issue) by Thompson, Knapp, and Levin. It is concluded that the controversy over statistical significance testing will no doubt continue. Nevertheless, the gradual movement of the field toward requiring additional information in the reporting of statistical results is viewed as evidence of a positive response to long-term criticisms of statistical significance testing.

In this rejoinder, I would like to (a) respond to the critiques of Bruce Thompson, Tom Knapp, and Joel Levin of my earlier article in this issue and (b) provide additional commentary as to the future direction of statistical significance testing.

Response to Three Critics

I would like to express my appreciation to the three respondents for their insightful observations and for their comments casting further light on the issues raised by the authors of the three articles appearing in this issue of the journal. Each of the respondents is a premier scholar whose contributions to the debates on statistical significance testing have been most useful as the issue has come to the forefront of methdological discussions in recent years. In their critiques of the three articles included in this issue, the three respondents have offered very useful discussions of the topic along with helpful references for those readers who might wish to explore the controversy further. My specific comments in relation to the points made by each respondent follow in the order in which they appear in this issue of the journal.

Bruce Thompson (1998) provides a nice framework for understanding the ongoing dialogue regarding statistical significance testing. Thompson's reminder of the context of the current literature in which much of the controversy has developed is useful in understanding the issue. This serves as a good follow up to the historical perspective that I provided. As Thompson noted, I have shared a long association with him and his work (he has been a mentor, research collaborator, and fellow editor); hence, I was not surprised that he was in agreement with many of the points I had raised and that a number of the opinions he expressed were consistent with my own. Further, I appreciate his citing the newly revamped editorial policies of several journals in addition to those that I had mentioned, lending evidence to the importance of editorial policies in shaping practice related to the reporting of results of statistical significance tests (SSTs). Further, Thompson (1998) reiterated nicely my discussion on the inappropriateness of using SSTs for the reporting of nil hypotheses about validity and reliability coefficients.

I am sure that Tom Knapp (1998) anticipated that the other authors and I would be eager to respond to his list of our various "errors of commission and omission." Obviously, determining what constitutes a sin is at least somewhat dependent upon the particular book of faith to which one prescribes. Although I prefer a slightly different statistician's book of faith than the one Knapp uses, I would have to say I am guilty as charged on at least a few points. First, I appreciate Knapp's (1998) comment on the distinction between the obtained and hypothesized effect sizes, an issue that often gets lost in the discussions of issues of this type. Second, I did indeed omit Levin's (1998a) excellent review of the *What If* book (Harlow, Mulaik, & Steiger, 1997) from my original discussion. This review is noteworthy not only because of Levin's excellent review of the content of the various chapters of the book, but also due to the concise list of recommended statistical significance practices that Levin offers. Third, I did not specifically mention the chapter in the *What If* book by Abelson (1997), which as Knapp (1998) indicated, is one of the more tightly written defenses of statistical significance testing.

Now that I have duly confessed, I would like to make a few citations from my own statistical book of faith on a couple of Knapp's other points. First, Knapp (1998) commented that resampling techniques such as jackknife and bootstrap analyses do not provide evidence of result replicability. (Levin [1998b] levels somewhat different but similarly focused criticisms at these procedures.) Even though the developers of jackknife and bootstrap techniques may not have specifically mentioned the usefulness of these procedures in providing evidence of replicability, the procedures do indeed create varied resamplings for which results may be recomputed many times over. Clearly, the replications of results from these resamplings are somewhat biased and do not replace

---

actual replications of the results with independent samples, but in newer areas of research, biased estimates of result replication are definitely better than no estimates of replication at all.

Knapp (1998) also questions the usefulness of "what if" analyses in which the results of SSTs are referenced to variations in sample size. Although I appreciate Knapp's concern that sample size should be carefully considered prior to the initiation of a study, it is often useful to determine at what sample size a statistically significant result would have become statistically nonsignificant and at what point a statistically nonsignificant result would have become statistically significant. These findings may advise researchers in selecting samples for *future* studies.

Knapp (1998) also splits hairs over the definition of the null hypothesis, apparently hinting at Cohen's distinction between null hypotheses in their most "general sense" and "the nil hypothesis" that states that "the effect size (ES) is 0" (Cohen, 1994, p. 1000). Although this is an important distinction, Cohen (1994) reminded us that "as almost universally used, the null in H_0 is taken to mean nil, zero" (p. 1000); hence, my use of this conventional definition. Similarly, Knapp (as well as Levin, 1998b), commented on the technicalities of my example comparing SSTs with an *n* of 62 versus an *n* of 302. My intent was not to suggest that the relationship between *p* and F is linear, but rather to show with a fixed effect that results that were not statistically significant given a particular sample size would be much more likely to be statistically significant given a larger sample size.

Levin (1998b), in his predictably amusing style, provided some excellent comments on the several papers and the controversy. His comments on "statistical testiness" are especially interesting. As Thompson (1998) noted, not all scholars will have totally positive opinions about editorial policies, such as the ones I prescribed, that encourage specific practices in the reporting of the results of SSTs. Here, Levin voices at least one oft-heard complaint leveled at such editorial policies, namely, that regulation of specific verbiage transforms editors from being scholarly gatekeepers to statistical police. Although I am an ardent supporter of academic freedom, I do feel that regulation of vocabulary so as to avoid miscommunication is essential, and, as an editor, I have with some frequency felt it necessary to correct authors' verbiage so as to enhance their clarity of communication. Without a doubt, the term "significant" constitutes one of the more significant (pun intended) instances of miscommunication in social science literature, especially among readers who may not be familiar with the logic underlying SSTs. And, even though, as Levin (1998b) suggested, the specific written context may sometimes disambiguate the use of the term "significant," I would prefer to require routine use of "statistically" before "significant" so as to avoid overlooking instances in which the term should have been modified thusly but was not.

I feel that Levin somewhat overstated my position on statistical significance testing when he suggested I advocated that "the research world will be a far better place when the hypothesis-testing devil is ousted by the effect-size angel." Although I would clearly acknowledge the heavenliness of effect size reporting, I do not see hypothesis testing as the devil, but rather as an oft-tormented, though well-intended, soul who needs the demon of misinterpretation exorcized from him. In fact, in this regard, my position is not extremely dislike the one stated by Levin: report both effect size estimates and results of SSTs, then allow the readers of the research report to draw their own conclusions about result importance.

Comments on the Future of Statistical Significance Testing

Contrary to Levin's hopeful assertion that perhaps one day soon the bickering over statistical significance testing will be quelled, I do not see that happening very soon. Rather, I agree with Thompson (1998) that the status quo regarding the use of statistical significance testing is far from "peachy keen." Unfortunately, the literature is still rife with studies in which authors have misused and misinterpreted SSTs. As long as this remains the case, the voices of reformers as well as defenders of statistical significance testing will continue to be loudly heard. The battle will continue to rage for some time to come with perhaps an occasional quietus as other important methodological issues emerge followed by rekindling of the flames of debate as thoughtful researchers continue to see errors in the reporting of SSTs.

Despite the slowness of progress in reforming practice relative to statistical significance testing, it is encouraging to see that an increasing number of social science journals are adopting editorial policies that call for better reporting of the results of SSTs (Thompson, 1998) following the suggestions found in the APA manual (APA, 1994). The adoption and enforcement of stricter editorial policies regarding the reporting of the results of statistical significance testing by an increasing number of social science journals will perhaps eventually move the field toward improved practice. At the recent annual meeting of the Mid-South Educational Research Association, Jim McLean, Co-Editor of this journal held a session in which he solicited input from the association members regarding the journal's potential adoption of an editorial policy on statistical significance testing. As a session participant, I was pleased to see that the group overwhelmingly favored such a policy. I look forward to seeing how Jim and Co-Editor Alan Kaufman handle the input gathered during that session.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would have to be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington: Author.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Knapp, T. R. (1998). Comments on the statistical significance articles. *RESEARCH IN THE SCHOOLS*, *5*(2), 39-41.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement*, *58*, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *RESEARCH IN THE SCHOOLS*, *5*(2), 43-53.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS*, *5*(2), 33-38.

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 67-69

TITLE INDEX
Volumes 1 - 5

A Case Study of an In-School Suspension Program in a Rural High School Setting . . . 4(2),57-64
A Comparison of Two Procedures, the Mahalanobis Distance and the Andrews Pregibon Statistic, for Identifying Multivariate Outliers . . . 1(1), 49-58
A Review of Hypothesis Testing Revisited: A Rejoinder to Thompson, Knapp, and Levin (by Nix and Barnette) . . . 5(2), 55-57
A Survey of Accelerated Master of Teaching Program Graduates at The University of Memphis . . . 3(1), 61-66
A Typology of School Climate Reflecting Teacher Participation: A Q-technique Study . . . 2(2), 51-57
An Analysis of the Charles F. Kettering Climate Profile . . . 1(2), 37-46
Applications, Trials, and Successes of CU-SeeMe in K-12 Classrooms in the Southeast . . . 5(1), 11-17
Are Students Overly Confident in Their Mathematical Errors? . . . 3(2), 1-8
Aspirations of Minority High School Seniors in Relation to Health Professions Career Objectives . . . 1(1), 21-28
Assessing School Work Culture . . . 4(1), 35-44
Assistant Principals' Concerns About Their Roles in the Inclusion Process . . . 4(1), 57-66
Association Between Metals and Cognitive, Psychological, and Psychomotor Performance: A Review of the Literature . . . 4(1), 1-8
Beam Me Up Scottie: Professors' and Students' Experiences With Distance Learning . . . 3(2), 35-40
Biology Students' Beliefs about Evolutionary Theory and Religion . . . 2(2), 31-38
Comments on the Statistical Significance Testing Articles . . . 5(2), 39-41
Differences in Reading and Math Achievement Test Scores for Students Experiencing Academic Difficulty . . . 3(2), 15-22
Do Funding Inequities Produce Educational Disparity? Research Issues in the Alabama Case . . . 1(2), 3-14
Dropping Out of Secondary School: A Descriptive Discriminant Analysis of Early Dropouts, Late Dropouts, Alternative Completers, and Stayins . . . 5(1), 1-10
Effective Teaching Behaviors for Beginning Teachers: A Multiple Perspective . . . 3(1), 1-12
Effects of Item Parameters on Ability Estimation in Item Response Theory . . . 1(2), 77-84
Effects of Learning Style Accommodation on Achievement of Second Graders . . . 3(2), 9-14
Evaluation of the Teaching Enhancements Affecting Minority Students(TEAMS) Program . . . 4(2), 9-16
Fight the Good Fight: A Response to Knapp, Levin, and Thompson . . . 5(2), 59-62
Fourth and Fifth Grade Students' Attitudes Toward Science: Science Motivation and Science Importance as a Function of Grade Level, Gender, and Race . . . 5(1), 27-32
Gender Differences in Achievement Scores on the Metropolitan Achievement Test-6 and the Stanford Achievement Test-8 . . . 1(1), 59-62
How Experienced Teachers Think About Their Teaching: Their Focus, Beliefs, and Types of Reflection . . . 4(2), 25-38
In Memoriam [of Ralph W. Tyler] . . . 1(2), 1-2
Influences on and Limitations of Classical Test Theory Reliability Estimates . . . 3(2), 61-74
Internalizing/Externalizing Symptomatology in Subtypes of Attention-Deficit Disorder . . . 2(1), 17-26
Leadership for Productive Schools . . . 1(1), 29-36
Lessons in the Field: Context and the Professional Development of University Participants in an Urban School Placement . . . 2(1), 41-54
Locus of Control, Social Interdependence, Academic Preparation, Age, Study Time, and Study Skills of College Students . . . 2(1), 55-62
Matching Reading Styles and Reading Instruction . . . 2(1), 11-16
Metaphor Analysis: An Alternative Approach for Identifying Preservice Teachers' Orientations . . . 1(2), 53-60

TITLE INDEX

Modeling Asymmetric Hypotheses with Log-Linear Techniques . . . . . 5(1), 61-68

Preservice Teachers and Standardized Test Administration: Their Behavioral Predictions Regarding Cheating . . . . . 2(2), 47-50

Preservice Teachers in Two Different Multicultural Field Programs: The Complex Influences of School Context . . . . . 3(2), 23-34

Preservice Teachers' Views on Standardized Testing Practices . . . . . 2(1), 35-40

Prevalence and Identification of Attention-Deficit Hyperactivity Disorder in a Mid-Southern State . . . 4(2), 49-56

Principal Leadership Style, Personality Type, and School Climate . . . . . 2(2), 39-46

Proper Use of the Two-Period Crossover Design When Practice Effects are Present . . . . . 4(1), 67-72

Quantitative Graphical Display Use in a Southern U.S. School System . . . . . 5(1), 33-42

Racial Identity Attitudes and School Performance Among African American High School Students: An Exploration Study . . . . . 4(2), 1-8

Reliability and Validity of Dimensions of Teacher Concern . . . . . 2(1), 27-34

Responses That May Indicate Nonattending Behaviors in Three Self-Administered Educational Surveys . . . . . 3(2), 49-60

Retention Across Elementary Schools in a Midwestern School District . . . . . 2(2), 15-22

School Counselors' Perceptions of the Counseling Needs of Biracial Children in an Urban Educational Setting . . . . . 4(2), 17-24

School Transformation Through Invitational Education . . . . . 2(2), 1-6

Score Comparisons of ACCUPLACER (Computer-Adaptive) and COMPANION (Paper) Reading Tests: Empirical Validation and School Policy . . . . . 4(2), 65-71

Secondary School Size and Achievement in Georgia Public Schools . . . . . 5(1), 18-26

Self-esteem and Achievement of At-risk Adolescent Black Males . . . . . 1(2), 23-28

Sequential-Simultaneous Profile Analysis of Korean Children's Performance on the Kaufman Assessment Battery for Children(K-ABC) . . . . . 1(2), 29-36

Sixty Years of Research in the Schools: A Conservation with Ralph W. Tyler . . . . . 1(1), 1-8

Small is Far Better . . . . . 1(1), 9-20

Staff Development for Improved Classroom Questioning and Learning . . . . . 2(1), 1-10

Statistical Significance and Effect Size Reporting: Portrait of a Possible Future . . . . . 5(2), 33-38

Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals . . . . . 5(2), 23-32

Statistics Anxiety: A Function of Learning Style? . . . . . 5(1), 43-52

Student Self-Concept-As-Learner: Does Invitational Education Make a Difference? . . . . . 1(2), 15-22

Students and the First Amendment: Has the Judicial Process Come Full Circle? . . . . . 1(2), 47-52

Support for Prayer in the Schools Among Appointed and Elected Alabama Superintendents and School Board Members . . . . . 4(1), 9-16

Suspensions of Students With and Without Disabilities: A Comparative Study . . . . . 4(1), 45-50

Teacher Perception of Kentucky Elementary Principal Leadership Effectiveness and School-Based Council Meeting Effectiveness . . . . . 4(2), 39-48

Testing at Higher Taxonomic Levels: Are We Jeopardizing Reliability by Increasing the Emphasis on Complexity? . . . . . 3(1), 45-50

The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing . . 5(2), 3-14

The Demise of The Georgia Teacher Performance Assessment Instrument . . . . . 3(2), 41-48

The Effect of Random Class Assignment on Elementary Students' Reading and Mathematics Achievement . . . . . 2(2), 7-14

The Effects of Specific Interventions on Preservice Teachers' Scores on the National Teacher Exam . 4(1), 51-56

The Effects of Violations of Data Set Assumptions When Using the Oneway, Fixed-Effects Analysis of Variance and the One Concomitant Analysis of Covariance . . . . . 1(2), 61-76

The Global Coherence Context in Educational Practice: A Comparison of Piecemeal and Whole-Theme Approaches to Learning and Teaching . . . . . 1(1), 63-76

TITLE INDEX

The Harrington-O'Shea Career Decision-Making System(CDM) and the Kaufman Adolescent and Adult Intelligence Test (KAIT): Relationship of Interest Scales to Fluid and Crystallized IQS at Ages 12 to 22 Years . . . . 2(1), 63-73
The Prediction of Academic Achievement Using Non-Academic Variables . . . . 3(1), 35-44
The Relationship of Student Attitudes Toward Science, Mathematics, English and Social Studies in U.S. Secondary Schools . . . . 3(1), 13-22
The Relationship of the Murphy-Meisgeier Type Indicator for Children to Sex, Race, and Fluid-Crystallized Intelligence on the KAIT at Ages 11 to 15 . . . . 1(1), 37-48
The Role of Statistical Hypothesis Testing in Educational Research . . . . 5(2), 15-22
The Statistical Significance Controversy is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin . . . . 5(2), 63-65
The Selection of Female Secondary School Assistant Principals and Transformational Leadership . . . 3(1), 51-60
The Verbal and Nonverbal Intelligence of American vs. French Children at Ages 6 to 16 Years . . . . . 3(1), 23-34
Topic Coverage in Statistics Courses: A National Delphi Study . . . . 5(1), 53-60
Using a Priori Versus Post-Hoc Assignment of a Concomitant Variable to Achieve Optimal Power from ANOVA, Block, and ANCOVA Designs . . . . 3(1), 67-82
Using Research Results on Class Size to Improve Pupil Achievement Outcomes . . . . 2(2), 23-30
What if There Were No More Bickering About Statistical Significance Tests? . . . . 5(2), 43-53
Written Expression Reviewed . . . . 4(1), 17-34

Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 71-72

AUTHOR INDEX
Volumes 1 - 5

Achilles, Charles M. . . . . 1(1), 9-20;
2(2), 7-14 & 23-30
Anderson , Robert H. 4 (1), 35-44
Arnold, Margery E. . . . . 3(2), 61-74
Austin, Sue . . . . 4(1), 51-56
Bailey, Tiffany L. . . . . 3(1), 61-66
Baldwin, Beatrice . . . . 2(2), 31-38
Barnette, J. Jackson . . . . 2(1), 1-10;
3(2), 49-60; 4(1), 9-16; 5(2), 3-14, 55-57
Barry, Nancy H. . . . . 3(1), 1-12
Bauschlicher, Lynn . . . . 1(1), 59-62
Bean, Rita M. . . . . 4(2), 25-38
Ben-Zeev, Talia . . . . 3(2), 1-8
Benson, William H. . . . . 4(1), 1-8
Billingsley, Collin . . . . 4(1), 1-8
Bogotch, Ira E. 2(2), 51-57
Bol, Linda . . . . 3(1), 61-66;
4(2), 17-24
Boyd-Zaharisas, J. . . . . 1(1), 9-20
Britt, Susan E. 3(1), 35-44
Burgess, Patricia D. . . . . 4(1), 57-66
Byun, Chang-Jin . . . . 1(2), 29-36
Cain, Van A. . . . . 2(2), 7-14
Campbell, Todd C. . . . . 5(1), 1-10
Carter, Richard B. . . . . 2(1), 35-40
Clements, Andrea D. . . . . 3(1), 45-50
Cobbs, Charles R. . . . . 1(2), 23-28
Cole, Jason C. . . . . 4(1), 17-34;
4(2), 65-70
Conners, Nicola A. . . . . 4(2), 9-16
Crawford, Melissa, ... 4(2), 9-16
Daley, Christine E. . . . . 4(2), 49-56
Daniel, Larry G. . . . . 5(2), 23-32,
63-65
DeMoulin, Donald F. . . . . 1(2), 47-52
Dempsey, John V. . . . . 5(1), 33-42
Denk, James P. 1(1), 21-28
Douzenis, Cordelia . . . . 1(2), 3-14
Downing, Hunter . . . . 4(1), 51-56
Duffy, Michael . . . . 5(1), 1-10
Egelson, Paula . . . . 2(2), 23-30
Eltinge, Elizabeth M. . . . . 5(1), 53-60
Enger, John M. 1(2), 23-28
Ernest, James M. . . . . 5(2), 15-22,
59-62
Fasko, Daniel . . . . 4(1), 45-50
Fisher III, Samual H. . . . . 5(1), 33-42
Frate, Dennis A. . . . . 4(1), 1-8
Fulmer, Deborah . . . . 4(2), 25-38
Fulton, B. D. . . . . 1(1), 9-20
Gipe, Joan P. . . . . 1(2), 53-60;
2(1), 41-54; 3(2), 23-34
Gonzalez, Jose J. . . . . 2(1), 17-26
Griffin, Harold . . . . 4(2), 49-56
Grubb, Deborah J. . . . . 4(1), 45-50
Grumet, Judith V. . . . . 4(2), 25-38
Gurewitz, Sim . . . . 2(2), 15-22
Gutkin, Terry B. . . . . 4(2), 1-8
Hale, Judith B. 5(1), 33-42
Haley, Kathleen A. . . . . 4(1), 17-34
Halpin, Gerald . . . . 3(2), 9-14;
5(1), 61-68
Halpin, Glennelle . . . . 3(2), 9-14;
5(1), 61-68
Hanson, Richard . . . . 4(2), 9-16
Hardin, Dawn T. . . . . 2(2), 39-46;
4(2), 57-64
Harman, Patrick . . . . 2(2), 23-30
Hassenpflug, Ann . . . . 3(1), 51-60
Howerton, D. Lynn . . . . 1(2), 23-28
Hynd, George W. . . . . 2(1), 17-26
Iran-Nejad, Asghar . . . . 1(1), 63-76
Ittenbach, Richard F. . . . . 4(1), 1-8
Jarrell, Michele Glankler . . . . 1(1), 49-58
Johnson, Annabel M. . . . . 1(1), 29-36;
1(2), 37-46; 4(1), 35-44
Johnson, Collen Cook . . . . 1(2), 61-76
Johnson, William L. . . . . 1(1), 29-36;
1(2), 37-46; 4(1), 35-44
Jones, Craig H. 1(1), 59-62;
2(1), 55-62; 3(2), 15-22 ; 5(1), 27-32
Jones, Randall . . . . 2(1), 35-40
Juergens, John P. . . . . 4(1), 1-8
Kaufman, Alan S. . . . . 1(1), 37-48;
1(2), 29-36; 2(1), 63-73; 3(1), 23-34
Kaufman, James C. . . . . 3(1), 23-34
Kaufman, Nadeen L. . . . . 3(1), 23-34
Kazelskis, Richard . . . . 2(1), 27-34
Keizer, Jenka . . . . 5(1), 11-17
Kher-Durlabhji, Neelam . . . . 2(1), 35-40;
3(2), 35-40
Kim, Jwa K. . . . . 1(2), 77-84;
3(1), 35-44
Knight, Carol Bugg . . . . 3(2), 9-14

AUTHOR INDEX

Kramer, Jack . . . . 2(2), 15-22
Lacina-Gifford, Lorna J. . . . . 2(1), 35-40; 3(2), 35-40
Lacour, Eileen . . . . 4(1), 51-56
Lacy, Elizabeth M. . . . . 4(1), 9-16
Lawrence, Frank . . . . 5(1), 61-68
Levin, Joel R. 5(2), 43-53
Lindauer, Patricia . . . . 4(2), 39-48
Lumpe, Andrew T. . . . . 3(1), 13-22
Lutkus, Anthony D. . . . . 4(2), 65-70
MacKay, Louise L. . . . . 4(1), 57-66
Marini, Irmo . . . . 2(1), 55-62
Martin, Ellice P. . . . . 5(1), 18-26
Martin, Nancy K. . . . . 2(2), 47-50; 4(1), 51-56
McFadden, A. C. . . . . 5(1), 11-17
McGinty, Dixie . . . . 3(2), 41-48
McLean, James E. . . . . 1(1), 1-8, 37-48; 1(2), 3-14, 29-36; 2(1), 63-73; 3(1), 67-82; 5(2), 15-22, 59-62
Miller, Leslie Michel . . . . 1(1), 21-28
Mittag, Kathleen Cage . . . . 5(1), 53-60
Moody, M. Suzanne . . . . 4(1), 67-72
Moon, Soo-Back . . . . 1(2), 29-36
Moore, Ramona C. . . . . 2(1), 41-54; 3(2), 23-34
Muenz, Tracy A. . . . . 4(1), 17-34
Naumann, Wendy C. . . . . 4(2), 1-8
Nishimura, Nancy J. . . . . 4(2), 17-24
Nix, Thomas W. . . . . 5(2), 3-14, 55-57
Nunnery, John . . . . 1(2), 3-14
Nye, B. A. . . . . 1(1) 9-20
Oglesby, Frankie . . . . 2(1), 11-16
Oliver, J. Steve . . . . 3(1), 13-22
Onwuegbuzie, Anthony J. . . . . 4(2), 49-56; 5(1), 43-52
Orletsky, Sandra . . . . 2(1), 1-10
Osborne, Jeanne S. . . . . 4(1), 45-50
Petrie, Garth . . . . 4(2), 39-48
Purkey, William Watson . . . . 1(2), 15-22; 2(2), 1-6
Rakow, Ernest A. . . . . 1(2), 61-76
Reeves, Carolyn K. . . . . 2(1), 27-34
Rice, Margaret L. . . . . 5(1), 11-17
Richards, Janet C. . . . . 1(2), 53-60; 2(1), 41-54; 3(2), 23-34
Richardson, Michael . . . . 4(2), 39-48
Ross, Steven M. . . . . 1(2), 3-14
Rothenberg, Lori . . . . 3(1), 45-50
Sandoval, Steve R. . . . . 4(2), 1-8
Saphore, Roger . . . . 5(1), 11-17
Sattes, Beth D. 2(1), 1-10
Schipull, Douglas W. . . . . 2(1), 27-34
Shannon, David M. . . . . 3(1), 1-12
Shargey, Bernice Ochoa . . . . 1(1), 21-28
Simon, Mireille . . . . 3(1), 23-34
Sinclair, Anne . . . . 2(2), 31-38
Slate, John R. . . . . 1(1), 59-62; 2(1), 55-62; 3(2), 15-22 ; 5(1), 18-26, 27-32
Smith, Lana J. 1(2), 3-14
Snyder, Karolyn J. . . . . 1(1), 29-36; 4(1), 35-44
Spencer, Rebecca M. . . . . 4(1), 1-8
Stanley, Paula Helen . . . . 1(2), 15-22
Strahan, David . . . . 2(2), 1-6
Suter, W. Newton . . . . 2(1), 11-16
Swetman, Daniel L. . . . . 3(1), 1-12
Taylor, Dianne L. . . . . 2(2), 51-57
Thompson, Bruce . . . . 1(1), 21-28; 2(2), 51-57; 5(2), 33-38
Thomson, William A. . . . . 1(1), 21-28
Trentham, Landa L. . . . . 1(2), 3-14
Turnbough, Rose . . . . 1(1), 59-62
Turpin, Tammye . . . . 4(2), 57-64
vonEschenback, John F. . . . . 3(1), 1-12
Wallenhorst, M. P. . . . . 1(1), 9-20
Walsh, Jackie A. . . . . 2(1), 1-10
Wang, Jianjun . . . . 3(1), 13-22
Wellhousen, Karyn . . . . 2(2), 47-50
Whiteside-Mansell, Leanne . . . . 4(2), 9-16
Wright, Vivian H. . . . . 5(1), 11-17
Wu, Yi-Cheng . . . . 3(1), 67-82
Zaharias, Jayne B. . . . . 2(2), 7-14
Zigmond, Naomi . . . . 4(2), 25-38

JOURNAL SUBSCRIPTION FORM

This form can be used to subscribe to RESEARCH IN THE SCHOOLS without becoming a member of the Mid-South Educational Research Association. It can be used by individuals and institutions.

◇ ◇ ◇

Please enter a subscription to RESEARCH IN THE SCHOOLS for:

Name: ______________________________

Institution: ______________________________

Address: ______________________________

____

			COST
Individual Subscription ($25 per year)	Number of years	______	________
Institutional Subscription ($30 per year)	Number of years	______	________
Foreign Surcharge ($25 per year, applies to both individual and institutional subscriptions)	Number of years	______	________
Back issues for Volumes 1, 2, 3, and 4 ($30 per Volume)	Number of Volumes	______	________
TOTAL COST:			________

MAKE CHECKS PAYABLE TO MSERA
SEND FORM AND CHECK TO:

Dr. James E. McLean, Co-Editor
RESEARCH IN THE SCHOOLS
University of Alabama at Birmingham
School of Education, 233 Educ. Bldg.
901 13th Street, South
Birmingham, AL 35294-1250

MID-SOUTH EDUCATIONAL RESEARCH ASSOCIATION

The Mid-South Educational Research Association (MSERA) was founded in order to encourage quality educational research in the Mid-South and to promote the application of quality educational research in schools. Members of MSERA share interests in educational research, development, and evaluation. While most members are from institutions of higher education, many others represent state departments of education, public and private agencies, and public school systems. Graduate students comprise a significant portion of the membership. A majority of MSERA members are from the six states represented by the organization, but others are from many other states and several foreign countries. The MSERA is the largest regional educational research organization in the country.

The organization provides several services for its members. The annual meeting, held every November, offers many formal and informal opportunities for professional development through special training courses, sharing of research findings, and programmatic interests with colleagues. Members receive a subscription to *RESEARCH IN THE SCHOOLS* and the *Mid-South Educational Researcher*. The MSERA also provides recognition and cash rewards for its outstanding paper, an outstanding dissertation, and professional service.

MSERA Membership/Renewal Form

(Please print or type)

Name ______________________________

Organization ______________________________

Address ______________________________

____

Telephone Work: ______________________

Home: ______________________

Fax: ______________________

e-mail: ______________________________

Amount Enclosed:	MSERA 1999 Membership ($25 professional, $15 student)	$______
	MSER Foundation Contribution	$______
	TOTAL	$______

Make check out to MSERA and mail to:

Dr. Robert Calvery
Southside School District
70 Scott Drive
Batesville, Arkansas 72501